UNIVERSITY OF TARTU

Faculty of Science and Technology

Institute of Technology

Gleb Kovalev

Potential of Artificial Genomes in Genome-wide Association Studies

Bachelor's thesis (12 ECTS)
Curriculum Science and Technology

Supervisor:

Burak Yelmen

Co-supervisor:

Kadir Aktas

Tartu 2021

Resümee/Abstract

Tehisgenoomide potentsiaal ülegenoomse assotsiatsiooniuuringus (GWAS)

Ülegenoomse assotsiatsiooniuuringu (inglise k genome-wide association study, lühend GWAS) pädevus tabada uusi haigusega-seotud-variante sõltub väga palju andmekogumi suurusest. Samas, liiga suur andmete hulk on teadlastele takistuseks kuna paljud genoomi andmebaasid ei ole andmekaitse tõttu vabalt kättesaadavad. Võimalik lahendus, genereerivad vastandvõrgustikud (inglise k generative adversarial networks, lühend GANs) on hiljuti näidanud võimekust luua tõepäraseid tehisinimgenoome (inglise k artificial human genome, lühend AG), mis anonüümselt asendaksid ligipääsmatuid andmeid. Käesolev uurimus kirjeldab ettevalmistavaid samme, et AG-de rakendust GWASis avastada. Eesti tüüp 2 diabeedi (T2D) andmete põhjal treeniti AG-de saamiseks iseseisvalt mudel juhtumi- ja kontrollgruppidega. Põhikomponendi analüüsi (inglise k principal component analysis, lühend PCA) tulemused määrasid ära, millal mudeli treenimine lõpetati. Arvutuslike piirangute tõttu lõigati treenimiseks genoomid 1000-ühe-nukleotiidi-polümorfismi-suurusteks osadeks. Saadud AG-d ühendati tagasi kokku terveteks kromosoomideks ja neid võrreldi päris genoomidega vastavalt populatsiooni struktuurile, kasutades nii PCA-d kui ka väikese alleelisageduse (inglise k minor allele frequency, lühend MAF) korrelatsiooni. Ka juhtumi- ja kontrollgrupi vahelisi suhteid analüüsiti eelnevalt mainitud meetodeid kasutades. Järgnevalt tehti GWAS-i nii Eesti andmete kui AG-dega. Töös avastati, et ühendatud AG-d grupeeruvad päris andmetega võrreldes erinevalt. Lisaks näidati, et AG juhtumi- ja kontrollgrupid on eristatavad pseudo-populatsiooni struktuurid. Peale selle leiti, et MAF uuringus on AG-de juhtumi- ja kontrollgruppide erinevus suurem kui päris genoomidel. Kokkuvõtteks jäi AG-de sooritus GWAS-is alla keskmise, näidates kõrgelt täispuhutud tulemusi ilmselt süstemaatiliste erinevuste tõttu MAF-i tulemustes juhtumi- ja kontrollgruppide vahel. Selles uuringus käsitleti mitut AG-de potentsiaalset takistust, mis tegutsevad anonümiseerivate proksidena GWAS-i rakendustes, lisaks pakume ka suuniseid edaspidisteks uuringuteks, soovitades alternatiivseid treeningmeetodeid ja võimalikke tehnilisi täiustusi.

CERCS: B110 Bioinformaatika, meditsiiniinformaatika, biomatemaatika, biomeetrika; B790 kliiniline geneetika

Märksõnad: Genoomika, populatsioonigeneetika, sügavad tehisnärvivõrgud, generatiivsed mudelid, ülegenoomse assotsiatsiooniuuringud (GWAS), tehisgenoomid

Potential of Artificial Genomes in Genome-wide Association Studies

The ability of genome-wide association studies (GWAS) to identify new disease-associated variants is highly dependent on sample size. At the same time, having a large enough dataset is currently a barrier for researchers, since many genomic databases are not freely accessible due to privacy concerns. As a potential solution to this problem, generative adversarial networks (GANs) have recently demonstrated the ability to create realistic synthetic artificial human genomes (AGs), which could serve as anonymous surrogates for inaccessible data. This study describes the preliminary steps towards exploring the possible applicability of AGs in GWAS. Using Estonian type 2 diabetes (T2D) data, AGs were generated by training the model independently on case and control groups, using coherent principal component analysis (PCA) results as a stopping criterion. Due to computational limitations, genomes were split into 1,000 SNPs chunks for the training. Obtained AGs were stitched back to full chromosomes and compared to real genomes based on population structure, estimated via PCA, and minor allele frequency (MAF) correlation. Additionally, relationships between case and control groups were assessed using the same methods. Subsequently, GWAS was conducted on both Estonian data and AGs. It was discovered that stitched AGs cluster differently from real genomes. Besides, we showed that AG cases and controls represent distinct pseudo-population structures. Furthermore, for AGs, differences in MAFs between cases and controls were greater than for real genomes. Eventually, AGs performed poorly in GWAS, showing highly inflated results, possibly due to the systematic differences in MAFs between case and control groups. In this study, we address several potential barriers for AGs serving as anonymous proxies in GWAS applications and provide directions for future research, suggesting alternative training approaches and potential technical improvements.

CERCS: B110 Bioinformatics, medical informatics, biomathematics, biometrics; B790 Clinical genetics

Keywords: Genomics, population genetics, deep neural networks, generative models, genome-wide association studies (GWAS), artificial genomes

Contents

List of Figures

List of Tables

Terms, abbreviations and notations

GWAS - Genome-wide association study

AG - Artificial genome

ANN - Artificial neural networks

PCA - Principal component analysis

T2D - Type 2 diabetes

DL - Deep learning

ML - Machine learning

SNP - Single nucleotide polymorphism

GAN - Generative adversarial network

MAF - Minor allele frequency

LD - Linkage Disequilibrium

NN - Neural Networks

PS - Population Stratification

SDG - Stochastic Gradient Descent

GDR - Generalized Delta Rule

HWE - Hardy-Weinberg Equilibrium

Introduction

Many common diseases, including cancer, cardiovascular and neurological diseases, and diabetes, are complex (or polygenic) traits. They are influenced by a combination of several genes and environmental factors, creating a challenge for understanding underlying mechanisms and providing personalized treatment. Consequently, there is a higher chance of poor therapeutic outcomes and negative side effects. As a result, an increasing number of researchers strive to uncover genetic factors that underpin disease development. [1,2] Up to date, genome-wide association studies (GWAS) are a major method for analyzing complex diseases since this approach allows to reveal genetic variants associated with a phenotypic trait without any prior knowledge on these variants. GWAS have unveiled new gene sets correlated to many diseases and offered multiple insights into their underlying molecular mechanisms. [1, 2, 6]

Following multiple studies, it has been demonstrated that a major factor driving novel discoveries in GWAS is the number of data samples. A good dataset is a foundation for many genetic studies, and larger sample sizes lead to greater statistical power and, as a result, better outcomes and new findings. [2, 3, 7, 8] At the same time, despite the constant growth of genomic databases over the last decade, a dataset of high quality and sufficient size remains a real limiting factor for researchers. Many genomic datasets are either not publicly accessible or require time-consuming and tedious application procedures. This problem could be potentially overcome with the help of a subclass of machine learning algorithms called generative models. [5]

Nowadays, we observe a tremendous influence of machine learning in different industries and research areas as these technologies advance. Deep generative models, in particular, have shown a remarkable ability to produce new realistic data samples, which can be used to augment, for instance, image and video databases [9]. Innovative machine learning technologies are extending to a range of applications and now gaining momentum in the field of genomics [10]. Recently, deep generative models were implemented as a novel approach to construct artificial human genomes (AGs), which have the potential to become anonymous data substitutes, simplify data access, and be used as augmentations tools to genetic datasets [5]. Given the nature of synthetically created data, the perspective of AGs in genomic studies such as GWAS remains to be explored.

Further research is necessary to understand whether human AGs created with generative adversarial networks (GANs) can be used in GWAS and to what extent GWAS findings from real genomes and AGs are comparable. In this study, we will take the first steps towards investigating AGs' potential to serve as anonymous surrogates for GWAS.

1 Literature review

In the present study, we use a combination of methods and concepts from various disciplines to investigate the feasibility of using artificial human genomes in genome-wide association studies. Therefore, it is better to review the literature for each of these fields separately. The "Genome-wide association studies (GWAS)" section 1.1 provides a definition of this methodology, key principles of conducting a study, and a brief overview of its important applications. It also highlights how GWAS could benefit from genomic data availability. A GWAS on type 2 diabetes (T2D) data will be conducted in this study. We use T2D as a use-case example since it is a widely studied binary (case/control) trait. Although we do not aim to identify novel variants, it would be reasonable to provide some background on this complex disease in the "Type 2 Diabetes (T2D)" section 1.2. In the "Machine learning (ML) and deep learning (DL)" 1.3 and "Generative models and generative adversarial networks (GANs)" 1.4 sections, we explain some basic concepts of ML and generative models, particularly focusing on the GAN architecture and working principles. In the "ML in genetics" section 1.5 we briefly discuss a new paradigm appearing in genomic analyses, also highlighting the importance of more open data sharing in genomics. Finally, in the "Artificial human genomes (AGs)" section 1.6, we explain the concepts behind the main focus of this study, AGs, covering their properties and potential applications.

1.1 Genome-wide association studies (GWAS)

Genome-wide association studies (GWAS) are an increasingly popular phenotype-first, non-candidate-driven approach in genetic research, which aims to identify associations between single nucleotide polymorphisms (SNPs), which drive genetic variation between individuals in a population, and phenotypic traits (Figure 1.1). The discovery of trait-associated SNPs may subsequently lead to new insights into biological mechanisms that underpin these phenotypes. GWA studies are particularly relevant for complex diseases such as diabetes which, unlike Mendelian disorders, are caused by multiple genetic variants as well as environmental factors. [6]

GWAS process is commonly divided into several stages, such as sample collection and phenotype determination, genotyping, quality control (QC), statistical analysis, and validation study. Individuals are divided into groups, according to their clinical manifestation. In the case of dichotomous disease phenotypes, we refer to these groups as disease cases and healthy controls. [6]

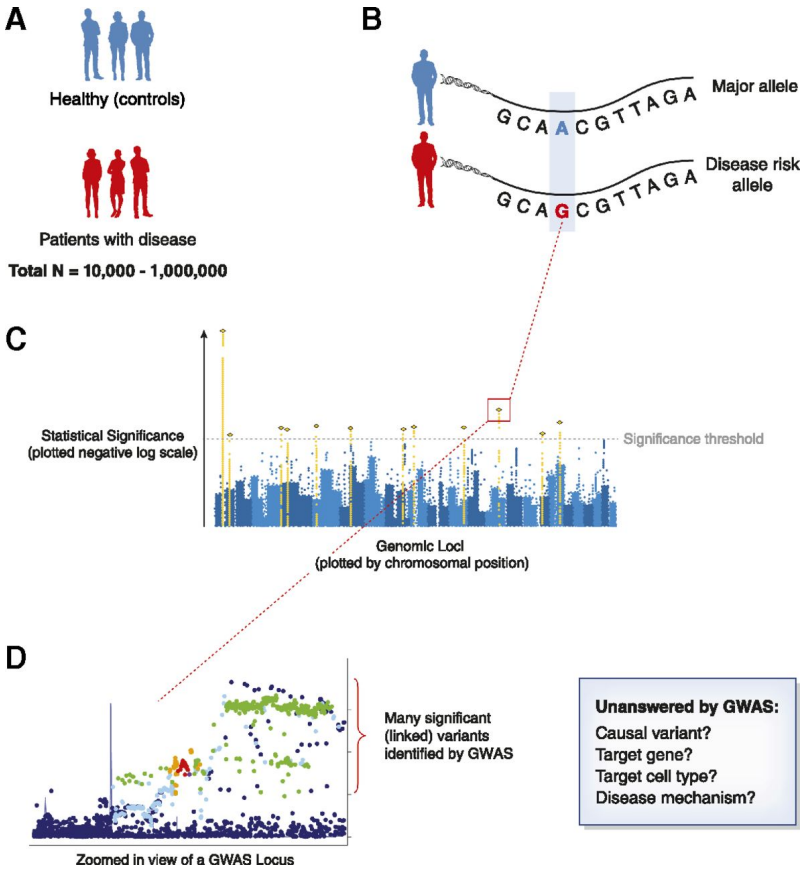

Figure 1.1: The Principle of a Genome-wide Association Study (GWAS) by Lin & Susztak (2020) [1]. In binary trait GWAS, individuals are usually divided into two groups: disease cases and healthy controls (A). Subsequently, single nucleotide polymorphism frequency is compared between these groups (B). Results are often displayed in the form of the so-called "Manhattan plot", where the most statistically significant SNPs rise like skyscrapers (C). Within one locus, many variants may indicate a correlation with disease development, as demonstrated in a locus zoom plot (D). It is important to note that the causal variant, the target gene, the target cell type, and the disease mechanism are not discovered from the GWAS.

GWAS is a genome-wide study of genotypes that can be collected through different technologies, such as whole-genome sequencing (WGS) and genome-wide SNP arrays. Nowadays, the majority of GWAS still use data from SNP arrays, although with technological advances and decreased costs future GWAS studies might rely more on WGS. SNP arrays do not capture all genetic variants in a population, thus the technology is usually combined with statistical imputation of missing genotypes from population reference panels. [2, 7]

Linkage disequilibrium (LD), a measure of non-random associations between DNA variants at distinct loci at the same chromosome in a given population, plays an important role in GWAS [6, 7]. Since multiple SNPs are inherited together, genotyping one or a few SNPs from each independent LD block is enough, whereas the rest of the variants can be inferred based on their LD [1]. SNPs that are in LD and inherited together are written in the form of haplotypes [2].

Appropriate QC is a vital step of any GWAS since raw genotype data is inherently imperfect for numerous reasons and will lead to unreliable results. Generally, during QC, SNPs and individuals are filtered out based on the individual and SNP missingness, sex discrepancy, minor

allele frequency (MAF), deviations from Hardy-Weinberg equilibrium (HWE), heterozygosity rate, relatedness, and population stratification. [6]

MAF is an important factor in GWAS since it partially affects the statistical power of detecting a significant correlation between variant and trait. A lower MAF for an allele (for example, a rarer variant) makes identification of an association less possible unless the variant has a significant phenotypic impact [1]. It is important to mention that the MAF of an SNP can differ between ethnicities with different evolutionary histories [11, 12], correlating with distinctions in disease prevalence between populations [1].

Common variants are thought to be evolutionarily old and shared among ethnic groups [2]. Nevertheless, since ancestry-driven allele frequency variations between populations may lead to biased results, it is critical to account for population ancestry before performing genotype–disease association analysis [13], emphasizing the significance of cohort demographics when planning, conducting and analyzing GWAS [1]. Genomic control analysis, structured analyses, and multivariate reduction analyses, such as principal component analysis (PCA) or multidimensional scaling (MDS), are commonly used methods to prevent the impact of population stratification on a study [14]. Usually, individuals beyond ethnicity borders are explicitly excluded from GWA studies on the basis of standard deviation units in principal-component dimensions [7]. As sample sizes get greater, however, it would be possible to not only use but also benefit from mixed and admixed ethnicity. For instance, the fact that allele frequencies and LD composition vary across populations can help in fine-mapping causal variants [7].

After performing QC, the data is prepared for subsequent association tests, which are appropriately selected based on the expected genetic model and nature of the phenotypic trait studied. Traits can be binary (e.g., disease cases and healthy controls) or quantitative (e.g., numeric, like a body-mass index) [6]. Generally, a large number of statistical tests are performed in parallel, each SNP being individually tested for association. The standard approach consists of computing individual, SNP-specific p-values corresponding to a statistical association test and comparing these p-values against some given significance threshold. SNPs with p-values below the threshold are considered as associated with a trait [15].

Since usually there is a large number of tests conducted, a considerable multiple testing burden is expected. Therefore, there is a need for multiple testing corrections. Various studies have revealed that the widely used genome-wide significance threshold of 5×10^{-8} for GWAS conducted on European populations adequately controls for the number of independent SNPs in the entire genome, regardless of the actual SNP density of the study. There are other common alternatives for determining genome-wide significance, including Bonferroni correction, Benjamini-Hochberg false discovery rate (FDR), and permutation testing. [6]

It is important to note that GWAS does not often identify causal variants and genes. Statistical strength of associations is not proportional to the biological significance of GWAS findings [2]. If the causal variant is in high LD with the tag-SNP, a correlation between a tag-SNP and a studied trait or disease can be indirect [1]. Moreover, the majority of association signals map to non-coding areas of the genome, which are notoriously difficult to biologically interpret [16,17]. As a result, after obtaining GWAS summary statistics, additional steps known as "post-GWAS" analyses are often taken to determine the causal variants and their target genes [1]. Several computational approaches such as fine-mapping, expression quantitative trait locus (eQTL) mapping, or functional annotation using epigenetic data, can be implemented to reveal a causal

variant underpinning an association signal [11].

If causal variants are successfully identified, a further guided functional investigation of genes and pathways responsible for a disease becomes possible. This new knowledge, in turn, could hold the key to better therapeutics for complex diseases. All this makes GWAS an important tool for disease diagnosis, drug discovery, and personalized medicine. [18]

The scientometric analysis of 3,639 GWAS studies from 2005 to 2018 demonstrated significant increases in sample sizes, rates of discovery and traits studied [8]. Although GWAS studies have scaled up to discover more variants, the full potential of GWAS is yet to be realized. Tam et al. (2019) described GWAS discoveries published to date as the "tip of the iceberg". The underwater portion of the iceberg represents the considerable number of findings that may be potentially discovered by including a broader spectrum of phenotypes, more diverse cultures and ethnic groups, different study designs, and considerably greater sample sizes. The sample size is considered to be the key limiting factor in risk variant finding. [2] Large sample sizes (e.g., at least in the order of thousands but likely even tens or hundreds of thousands) are necessary to detect genetic risk factors of complex traits. It is strongly recommended to avoid performing underpowered studies with a small number of samples. [6]

What is more, in complex trait GWAS, there is a sample size threshold above which the rate of locus discovery accelerates for each trait, and no trait up to date has shown signs of a plateau in the number of risk loci discovered as the sample size grows (Figure 1.2) [2, 7]. A good example of the "the more the merrier" principle is GWAS studies of Schizophrenia, where greater sample sizes led to more statistical power and new biological insights into this neuropsychic disorder [7]. Besides, very large sample sizes may be required to detect important gene-gene interactions [2].

In order to increase the power of studies and analyze millions of variants, modern GWAS combines data across multiple data sets in the form of meta-analysis. With a total of 33.71 authors per paper returned from the PubMed website, GWAS meta-analysis has historically required a partnership with multiple authors sharing data sets or expertise. [8]

Genotype and phenotype information on a large number of participants has been gathered, and in some cases is still being collected, by broad-scale programs in both the private and public sectors [2]. Private data in genotype-phenotype databases are strictly secured and therefore cannot be easily accessed. However, a significant number of studies recommend making use of resources for the good of more people and encouraging more open data-sharing. Experts believe that the public availability of genetic data and the possibility of combining different datasets will be a treasure trove for new fundamental human genetic discoveries [2, 3, 7, 8].

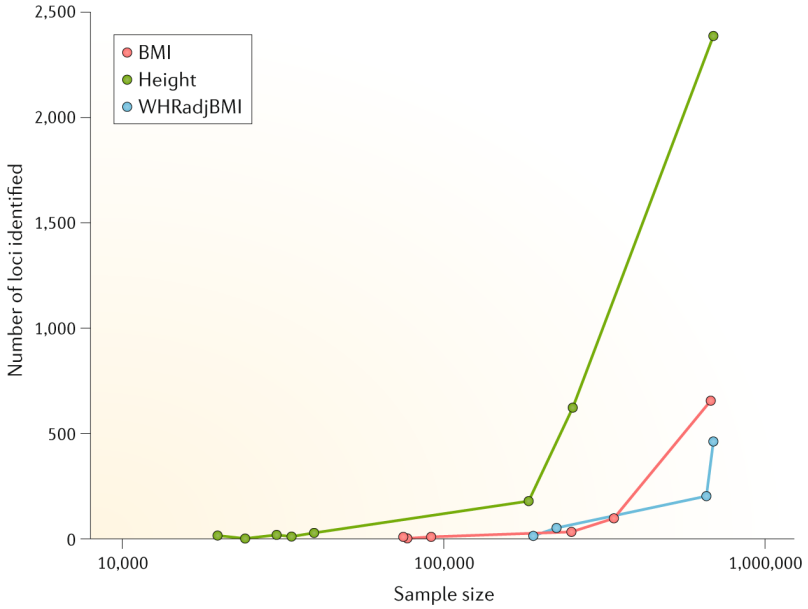

Figure 1.2: Number of loci identified as a function of GWAS sample size by Tam et al. (2019) [2]. The plot demonstrates the number of genome-wide significant loci ($P < 5 \times 10^{-8}$) recorded in GWAS in European or predominantly European populations for three anthropomorphic traits: body mass index (BMI), height, and waist-to-hip ratio adjusted for BMI (WHRadjBMI). After a certain inflection point in sample size, the number of identified loci exponentially increases.

1.2 Type 2 diabetes (T2D)

Type 2 diabetes mellitus (T2D) is a common heritable metabolic disorder with a high prevalence that illustrates many of the complexities and approaches for other complex diseases [3]. It is characterized by relative insulin deficiency caused by pancreatic-cell dysfunction, as well as insulin tolerance in target organs [19].

T2D accounts for more than 90% of diabetic patients and causes microvascular and macrovascular problems that inflict significant psychological and physical harm in both patients and caregivers, as well as a substantial financial burden on healthcare institutions [19]. The rising tide of obesity, sedentary lifestyles, and energy-dense diets has resulted in an exponential rise in the number of people diagnosed with T2D, which is expected to reach 642 million by 2040 [20].

Increased knowledge of particular diabetes phenotypes and genotypes may lead to more specific and tailored therapeutic approaches, allowing patients to be managed more efficiently [19]. While investigations of biological pathways for T2D are largely context-dependent, common methods have emerged in recent years [3]. A conventional "forward genetics" workflow involves conducting a GWAS of T2D case and control groups, identifying new loci, and then predicting causal variants via fine-mapping or expression quantitative trait loci (eQTLs) analysis [3].

The success of early GWAS for T2D was driven in large part by collaboration and data sharing among genetics researchers. Over time, T2D GWAS have grown in size and diversity, yielding more associated variants and offering key insights into T2D biology (Figure 1.3) [3]. It has been shown that the majority of risk variants span in non-coding regions of the genome and assist in

decreased β-cell function or mass [3], mostly affecting the regulatory mechanisms [21, 22].

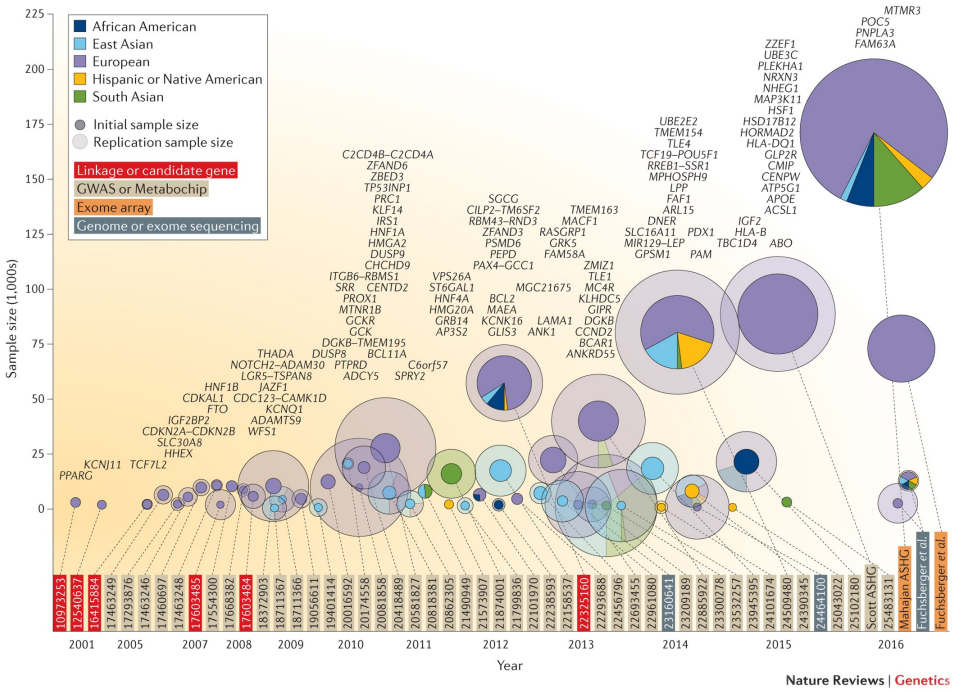

Figure 1.3: The history of T2D GWAS by Flannick & Florez (2016) [3]. The sample size and diversity in type 2 diabetes GWAS have been steadily growing over the years. In the figure, T2D GWAS, with additional candidate gene and sequencing studies, are plotted in the form of circles. The x-axis represents the year of publication, while the y-axis demonstrates the discovery sample size. The color of a circle is determined by the ethnic composition of the data set. Darker circles are scaled up in proportion to discovery sample size, while lighter circles are scaled in proportion to both discovery and replication sample size. PubMed identifiers or first author names for each study, colored according to the used technology, are indicated at the base of the figure. T2D associated loci are also annotated.

Despite the vast number of variants discovered by GWAS, the associated variants account for just a small portion of the heritability of T2D (about 10%). This is known as the "missing heritability" problem, which is likely caused by the presence of common variants ($MAF \geq 1\%$) that have small effects and have not been detected and/or rare variants ($MAF < 1\%$) that are not well tagged by common SNPs. [23]

To recognize or describe additional T2D associations, larger genetic studies will be required [3]. A recent meta-analysis study performed by Xue et al. (2018) [24] is a good example of a successful GWAS T2D analysis that highlights the benefits of very large sample size and the combination of multiple omics data. Researchers discovered 139 common and 4 rare variants associated with T2D, 42 of which (39 common and 3 rare variants) were not established previously. By integrating GWAS results with gene expression data from blood as well as DNA methylation and epigenomic annotation data, they also managed to identify and prioritize functional genes, proposing putative genetic regulatory mechanisms for T2D. What is more, the estimated genetic architecture suggests that T2D is a polygenic phenotype in which both common and rare variants contribute to the phenotype. It also implies that rarer variants have a tendency to have a greater impact on T2D risk.

Nowadays we can observe two trends in the research of T2D genetics. First, researchers believe that larger sample sizes will lead to the discovery of new disease-relevant variants. Second, as the number of genes or processes linked to T2D grows, so will the variety of approaches required to convert these associations into new knowledge of disease mechanisms. [3]

Flannick & Florez (2016) [3] argue that a new model for the transparent and collaborative exchange of data and findings between institutions would be most beneficial and synergistic for future research in each direction. The authors state that by "democratizing genetics" and allowing a wide number of users to carry out custom research, advancement in T2D biology research may accelerate, eventually leading to better care and outcomes for patients. They consider an integrated T2D knowledge base as one of the possible ways to optimize the global use of genetic data.

1.3 Machine learning (ML) and deep learning (DL)

Machine learning (ML) is a branch of artificial intelligence (AI) focused on the development of algorithms that automatically learn from a collection of data and improve their learning over iterations. While classical programming algorithms are coded based on known features, in ML algorithms are "trained" to find patterns and features in subsets of data, to improve algorithm accuracy rather than perform parameter estimation for a probabilistic model. Thus, even if our models are imprecise but adequate to reality, the ML approach can provide us with new insights into nature. [10, 25–27]

In ML, even though hybrid strategies exist, there are two major learning methods: supervised learning [27, 28] and unsupervised learning [27, 29] (Figure 1.4). Supervised learning uses patterns in the training dataset and maps input variables/features to the response variable/target, or label, which can be either categorical or continuous so that an algorithm can make predictions about new data points. The mapping is accomplished through learning the mapping function when a model is informed about the relationship between features and targets in the training dataset. The performance of the algorithm on the training dataset is compared to its performance on the validation dataset after each iteration of the training process, allowing the algorithm parameters to be fine-tuned. A loss function is a measure of the given prediction quality. The aggregated value of the loss function across the training set is defined as a risk function. In order to correctly predict the response variable, we want to minimize the risk function during the training process. Finally, the algorithm is evaluated on the test set that is independent of the training set. [10, 25–27, 30]

Regression and classification, depending on the response variable, are the most common supervised learning tasks. We refer to an ML task as a classification problem when the response variable is categorical. Classification involves predicting which category a data point belongs to. If the response variable is continuous, we define an ML task as regression, which entails continuous value prediction. [25–27, 30]

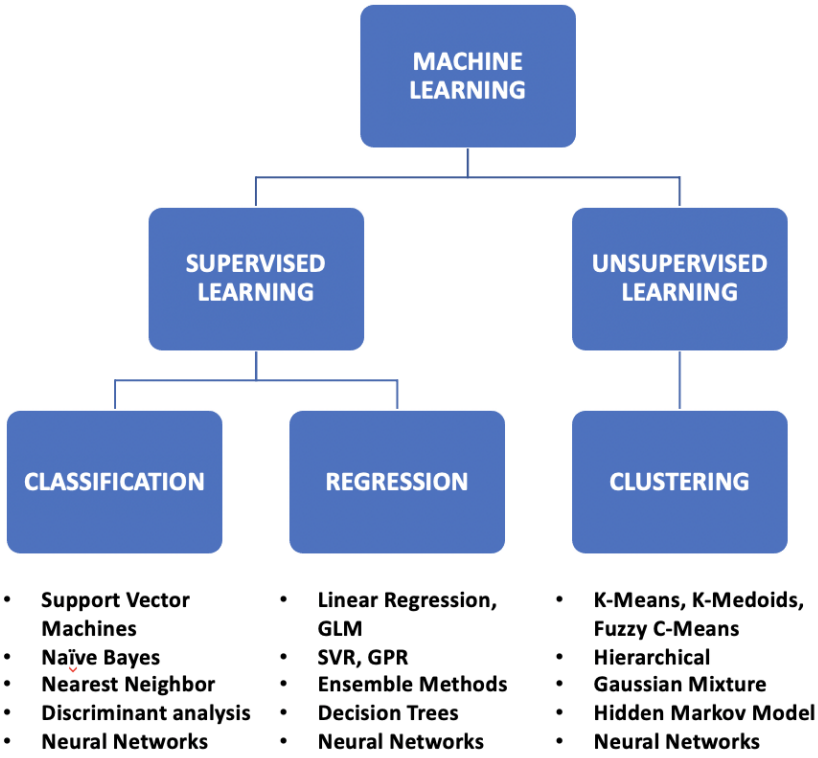

Figure 1.4: Common machine learning algorithms

Unsupervised learning, in comparison to supervised learning, seeks to find patterns and categorize individual instances within a dataset without knowing how the data is organized. Patterns that may or may not reside in the dataset are simply not defined by a target and should be determined by the algorithm. Clustering, association, and anomaly detection are some of the most general unsupervised learning task examples. [25–27, 30]

Deep learning (DL) is a subfield of ML focused on algorithms called artificial neural networks (ANN) or similarly networked algorithmic models that consist of multiple "hidden" layers between the input and output layers and are inspired by biological neural networks. These algorithms differ in their structure and training methods (Figure 1.5). Each ANN contains nodes (neurons) that communicate with one another via connections. Connections between nodes in an ANN are weighted based upon their capacity of providing the desired outcome. [25–27, 31, 32]

The basis of an ANN is the perceptron, which is a linear ML classifier algorithm that attempts to separate data points into classes in two-, three-, or hyper-dimensional space. Series of input features are transformed with the help of activation function, such as sigmoid function. [25, 27, 33, 34]

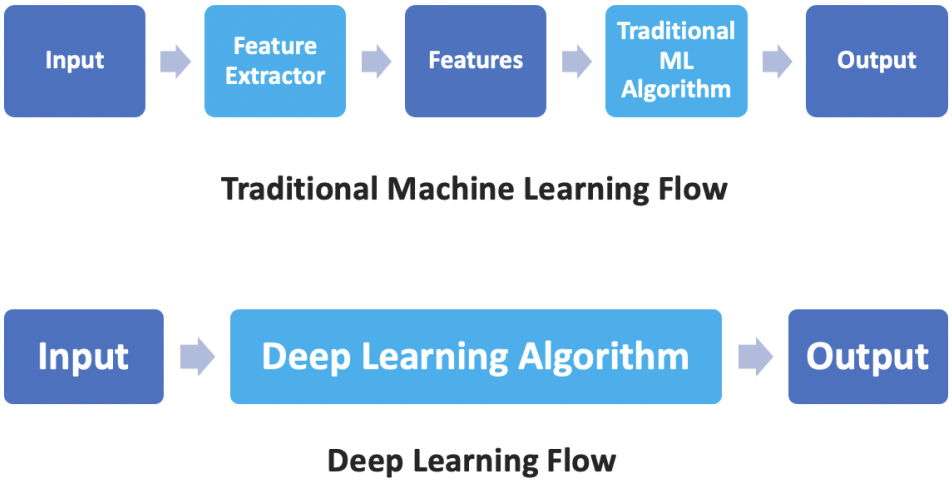

Figure 1.5: Differences between machine learning and deep learning flows

ANNs feed information forward for the majority of tasks when the information from each node in the previous layer is passed to each node in the next layer, transformed with activation function, and passed to the next layer. This is known as a feedforward neural network. [25, 27] For feedforward neural networks, backpropagation is a widely used training algorithm. Backpropagation utilizes the generalized delta rule (GDR) to compute the gradients for the network's parameters [27,35]. During the backpropagation, the network's weights are modified so that the error between the actual, desired output and the output predicted by ANN is minimized using a certain error function [36, 37]. In order to minimize an error function, optimization algorithms such as gradient descent are used. With gradients for the parameters of the neural network determined in backpropagation, during gradient descent, the model's parameters are gradually corrected in the steepest descent direction until the output error is minimized. The size of steps we take to get to the local minimum is determined by the learning rate constant [38]. The backpropagation approach includes two steps: propagating the input data through the network and working backward from the output layer to change the weights so that the average error across all layers is reduced [36, 37].

1.4 Generative models and generative adversarial networks (GANs)

The invention of new generative methods with the aim of producing synthetic data with ideal structures and properties is a big trend in deep learning (Figure 1.6). [39] Many ML classifiers such as support vector machine (SVM) or supervised feedforward deep artificial neural networks, focus on the discriminative classification process, where the decision boundary between the classes is modeled. Generative models, on the other hand, assume that the data is created by a certain probability distribution, which is then estimated, and a distribution very close to

the original one is generated. Both discriminative and generative classifiers have the same objective in mind: to calculate the probability of the target variable. However, discriminative models learn conditional probability, while generative models rely on finding joint probability by utilizing the Bayes theorem, which is far more informative and can be used to generate new data instances or select indicative features. In other words, discriminative models only draw a decision boundary in a data space, while generative models learn the overall distribution of the data. There are many examples of generative models that come from the traditional statistical approach as well as modern deep neural architectures. [4]

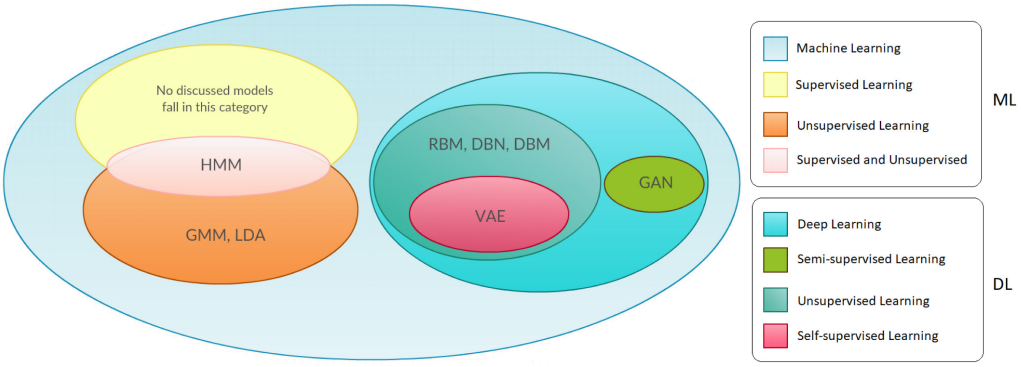

Figure 1.6: Classification of different generative models with respect to ML and DL by Harshvardhan et al. (2020) [4]

A general taxonomy of generative models, in the context of deep neural networks, was proposed by Goodfellow et al. (2014) [40]. Explicit density estimation models have an intractable explicit density function which requires approximation in order to maximize the likelihood. The difficulty with explicit density models is capturing all the complexities in a set of data while remaining tractable. Being unable to calculate the density itself, the implicit density models interact directly with a data distribution while training by sampling from it. In the case of generative adversarial networks (GANs), the sampling can be done directly, whereas, in the case of generative stochastic networks (GSNs), it can be done with the aid of a Markov chain. [4]

Generative models can be used when data is costly or inaccessible to simulate it. However, we will still need enough data to train a model. As an additional application, we may also utilize generative models as a tool for exploring new possible data configurations that extend the boundaries of our current understanding. [9]

In this literature review, our main focus will be directed to the generative adversarial networks (GANs). GANs are generative neural networks that were introduced by Goodfellow et al. (2014) [40]. In its basic form, the GAN model consists of two artificial neural networks which compete with each other in a zero-sum game. One of them, the generator, generates objects in the data space from some noise input (usually in the form of Gaussian or uniform distribution), and other, the discriminator, which is a multilayer perceptron, learns to distinguish objects generated by its partner from real examples from the training set. Thus, the network consists of two parts with opposite goals (Figure 1.7). [4, 9]

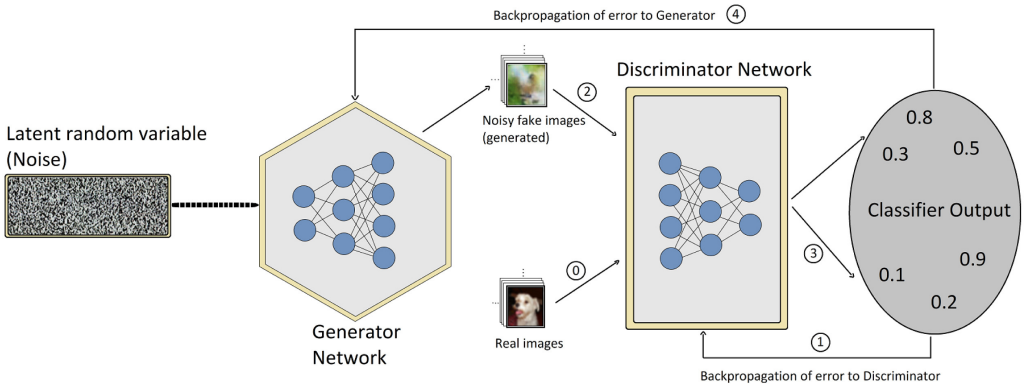

Figure 1.7: Generative adversarial network working principle by Harshvardhan et al. (2020) [4]. At first, real data is used to train a discriminator (0). The error is backpropagated through D, adjusting its weights (1). Subsequently, D is trained by feeding it by early-stage data samples produced by Generator (G), which takes noise as an input (2). D discriminates fake samples from real ones, returning values close to 0 (3). The output of D is subtracted by 1 and backpropagated through G, improving its ability to produce realistic samples (4). The steps are repeated for multiple epochs until D is unable to distinguish between generated and real data samples.

In other words, the discriminator solves the common problem of binary classification. Given an example, which looks like an element in a data space, the discriminator must decide whether it is a real one or was produced by the generator. [4, 9, 39] On the other hand, the generator aims to make it impossible for the discriminator to distinguish between the real data distribution and the distribution produced by the generator. Standard gradient descent algorithms can be used to train both the generator and the discriminator. When the discriminator learns to spot the difference between false and real data, it passes a valuable learning signal to the generator. [4, 9, 39]

Usually, the training process of GANs alternates between n steps of optimizing the discriminator and one step of optimizing the generator. The discriminator and the generator make each other better as they learn. Their minimax game terminates at the so-called Nash equilibrium [9]. After training is finished, we can use the generator to produce synthetic data [39].

The main advantage of GANs is that they allow training without having to go through the usually intractable process of optimizing log-likelihood, which typically requires numerous approximations. In contrast, traditional generative models assume that data follow a certain distribution and try to estimate it through maximum likelihood [4]. GANs do not require any prior assumptions and thus training is less complicated and more flexible. Besides, GANs do not require any Markov chains, which are cost-intensive, as in the case of Boltzmann machines [4]. Moreover, with an ability to capture the inherent rules of the natural world, GANs provide a new paradigm for unsupervised deep learning models [9].

GANs, on the other hand, have certain drawbacks, such as mode collapse, which occurs when a generator falls to the local statistical distribution mode, resulting in samples with low diversity. Another problem in GAN training is gradient disappearance, when a discriminator distinguishes between fake and real data too well, causing a generator's gradient to diminish. Furthermore, models may fail to converge due to uncontrollable training and parameter oscillations. [9]

Depending on different tasks and contexts, various GAN derivative models have been proposed

and extensively used in the field of computer vision. GANs also have demonstrated outstanding results in cross-domain areas such as medicine, art, and signal processing. What is more, research and possible implementations of GANs are considered to be in the relatively early stages, promising a wide range of applications and advances in the future. [4, 9, 39]

1.5 ML in genetics

While ML methods are powerful tools and have revolutionized data analysis in many fields, their application in population genetics interference is still at its dawn. To keep pace with constantly growing population genetic datasets, new computational methodologies are rapidly being developed by researchers in order to utilize genomic sequence data. The vast majority of population genetics research has concentrated on classical statistical estimation using a convenient probabilistic model, or a close approximation to that model. Such a model aims to adequately describe the data in a way that insights into nature can be gained through parameter estimation. [10]

Schrider & Kern (2018) [10] argue that researchers should consider utilizing ML approach as a powerful mode of analysis that has recently emerged within population genetics. In this paper, the authors describe several examples of how early applications of ML for population genetics can outperform traditional statistical approaches. They believe that supervised ML methods, which can take an advantage of high dimensional input, are valuable and underutilized tools with a lot of potential for evolutionary genomics.

Despite the fact that population genetic datasets are growing in size, nowadays, obtaining adequately sized datasets can be a challenging task due to the restricted accessibility by the governmental and private entities as well as privacy concerns [5]. For example, Estonian Biobank is not publicly available and requires approval from the Ethics Review Committee on Human Research of the University of Tartu as well as from the EGCUT scientific committee, provided that an applicant will send scientific results obtained from research conducted on the shared data [41]. What is more, many autochthonous populations are under-represented in genetics datasets, which limits the resolution of many studies, such as genome-wide association studies [42–45]. Besides, the number of developing machine learning algorithms in the field of genomics, including ones aiming to improve GWAS and post-GWAS analyses, are also often limited by the range and quality of training data [18].

Generative models may indeed have a potential solution to these problems. Although generative models have demonstrated impressive results in a broad range of domains, their capacity to generate meaningful synthetic data is still underutilized in genetics and until recently was absent from population genetics. [5]

Previously, Killoran et al. (2017) [39] investigated the ability of deep generative models to produce DNA sequences with certain biological properties. Results obtained by researchers indicate that generative models, including GAN, can learn the essential structure from DNA sequences, even from limited information, and be utilized for exploration and designing of new DNA sequences with desired properties. This work is the first detailed exploration of generative DNA design and provides initial validation of this methodology, opening a door to new research

direction with many use-cases in genetics. Moreover, based on their findings and extrapolating observations from the computer vision field, authors believe in the ability of GANs to scale up to large gene-scale sequences (thousands of nucleotides or more), enabling the design of entire genes and even small genomes in the future.

Interestingly, Schrider & Kern (2018) [10] were curious whether GANs can be used as a substitute for population genetic simulation and generate very large samples and chromosomes that are computationally costly to simulate.

1.6 Artificial human genomes (AGs)

For the first time, generative methods, particularly GANs and restricted Boltzmann machines (RBMs), were used in population genetics to create high-quality realistic Artificial Genomes (AGs) by Yelmen et al. (2021) [5].

In the research, various generative models were compared by their ability to create AGs. Generative models are evaluated based on the plausibility of generated samples relative to data distribution, mostly through visual inspection in the case of image generation, although it is difficult to unbiasedly assess produced data samples and further development is required to devise a better way of assessing the accuracy of generative models [4]. For AGs, researchers used principal component analysis (PCA) for the initial evaluation of models' performance (Figure 1.8). The PCA is an unsupervised machine learning method implemented in the context of population genetics for analyzing relatedness relationships among individuals. During the PCA, a high-dimensional genotype matrix is reduced to a lower-dimensional summary that shows genotype clusters [10]. Thus, by mapping generated artificial genomes to the real genomes, one can assess how close artificial genomes resemble the population structure of the real data.

Researchers demonstrated the ability of produced AGs to mimic the real genomic data and capture many complex characteristics such as allele frequencies, linkage disequilibrium, pairwise haplotype distances, as well as represent the underlying population structure of the real data, indicating overall promising applicability. They also showed that both RBM and GAN AGs can capture selection signals. Importantly, in both GAN and RBM cases, no real genome was copied into AGs. Researchers investigated whether models retained privacy by measuring the extent of overfitting and calculating two metrics of resemblance and privacy. As a result, AGs obtained with the GAN model demonstrated low privacy loss, while AGs obtained via RBM confirmed the presence of overfitting, which poses a high risk of privacy leakage. Since underfitting and low leakage information are preferred, GAN AGs have a slight advantage in this context.

AGs were also compared to advanced genome-generation methods and found to be advantageous in several cases. For instance, compared to the HAPGEN2 method which uses a copying model, significantly less overfitting and privacy loss were observed for AGs. For coalescent simulation, additional demographic parameters are required, and desired one-to-one SNP correspondence is not achievable, which prevents simulated genomes from being combined with real genomes for further analysis.

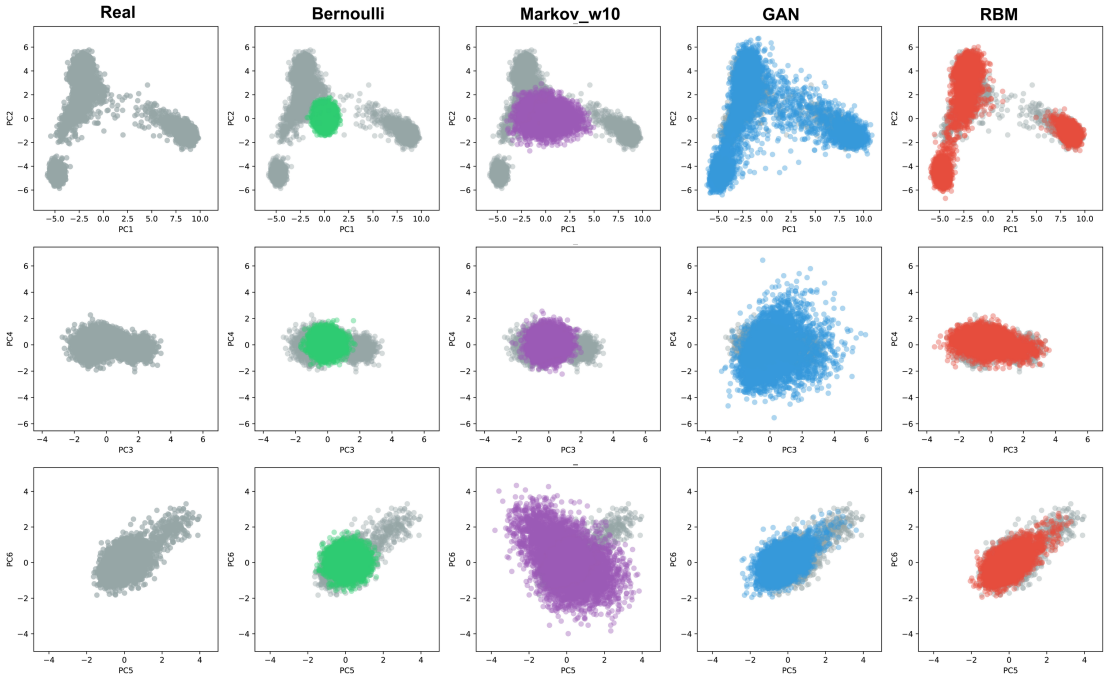

Figure 1.8: PCA analysis of real genomes with artificial genomes obtained via different generative models by Yelmen et al. (2021) [5]. The six first axes of a single PCA applied to real (gray) and artificial genomes (AGs) generated via Bernoulli (green), Markov chain (purple), GAN (blue) and RBM (red) models.

An important issue with the proposed GAN and RBM models is that, due to computational limitations, they can only be used to generate fragments or sequential dense chunks rather than whole artificial genomes. But, as the authors suggest, it should be possible to generate whole genomes by training and generating multiple chunks from different genomic regions independently using a single uniform population, such as Estonians, and stitching them together later to create genome-length sequences for each AG individual.

Another drawback comes from the inability of models to effectively capture rare alleles, especially for the GAN model. Mode collapse, which occurs when the generator fails to cover the full support of the data distribution, is a well-known problem in GAN training [46]. The observed inability of GANs to generate rare alleles may be due to this type of failure.

Authors believe that AGs can be used as alternatives for many genome datasets which are either not publicly available or require long application processes. Another possible application is to use the generative models' encoding of the real data as a starting input for multiple tasks, such as demography interference. Furthermore, by augmenting public genomic panels with AGs, it might be possible to enhance the performance of genomic tasks such as GWAS.

2 The aims of the thesis

The aims of the thesis are:

1. Conduct GWAS analysis on the Estonian T2D dataset;
2. Generate large artificial genome-like sequences based on the Estonian T2D data;
3. Conduct GWAS analysis on the obtained artificial genomes;
4. Perform a correlation analysis and compare the results of the two studies to assess whether artificial genomes can be utilized for GWAS.

Please see Figure 2.1 for the study design illustration.

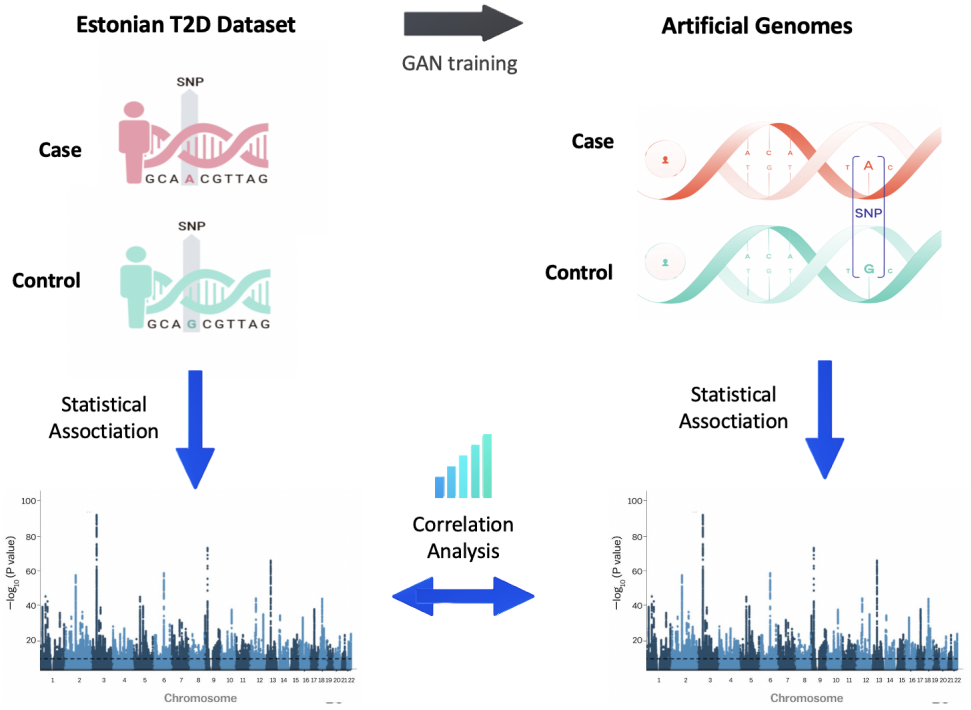

Figure 2.1: Study design. This research is divided into 4 parts. The first milestone is to perform GWAS analysis of Estonian Type 2 Diabetes data. The second milestone is to, based on the Estonian data, generate artificial genomes using the GAN model. The third milestone is to conduct GWAS analysis on the produced artificial genomes. Finally, the last step is to compare GWAS association results between real Estonian genomes and artificial genomes.

3 Experimental part

3.1 Materials and methods

3.1.1 Data

Data description

Genomes from Estonian Biobank were accessed with Approval Number 285/T-13 obtained on 17/09/2018 by the University of Tartu Ethics Committee. The Estonian Biobank [41] provided genomes (chromosomes 1-22) of 5,050 individuals, 2,525 of whom were diagnosed with type 2 diabetes (T2D) based on health records and the rest 2,525 were randomly selected as healthy controls. Available genotype data were based on the following genotyping technologies: whole-genome sequencing, Illumina HumanHap370CNV, Illumina HumanOmniExpress, Illumina HumanCoreExome, Illumina Global Screening Array. To make the generative adversarial network (GAN) model training feasible, Illumina HumanHap370CNV was used as a scaffold for downsampling single nucleotide polymorphism (SNP) sizes for all samples. Preliminary filtering was performed on the genotype data with VCFtools (0.1.15) [47]. Insertions and deletions were removed, only bi-allelic SNPs were kept. SNPs with minor allele frequency (MAF) < 0.01 were excluded. After this step, data consisted of 305,283 variants.

Quality control (QC)

As the next step, using PLINK version 1.9 [48], the genotype data were cleaned with the standard quality control (QC) steps. Individuals and SNPs with high levels of missingness were filtered, starting with a relaxed threshold of 0.2, followed by a more stringent threshold of 0.02. SNPs with $MAF < 0.01$ and Hardy-Weinberg equilibrium (HWE) test $P\text{-}value \leq 1e^{-10}$ in cases and $\leq 1e^{-6}$ in controls were excluded. Heterozygosity check and cryptic relatedness analysis were conducted on a set of pruned SNPs. Linkage Disequilibrium (LD) pruning was done using a multiple correlation coefficient of 0.2. Individuals with a heterozygosity rate deviating more than 3 standard deviations from the mean as well as related individuals at a genetic relatedness threshold of 0.125 (i.e., third-degree relatives) were excluded. As a result of QC, 4,452 individuals (2,024 cases and 2,428 controls) and 305,250 variants were retained for further analysis. Please see Table 3.1 for details.

Table 3.1: Quality control (QC) on the Estonian type 2 diabetes (T2D) data

Quality Control (QC)		
Step	**Thresholds**	**Result**
Missingness of individuals and SNPs	First, relaxed threshold: 0.2 (>20%); Second, more stringent threshold: 0.02 (>2%).	No SNPs and individuals are removed
Minor Allele Frequency (MAF)	0.01 (>1%)	28 SNPs are removed
Hardy-Weinberg equilibrium (HWE)	HWE p-value < 1e-10 in cases; HWE p-value < 1e-6 in controls.	5 SNPs are removed
Heterozygozity	± 3 SD from the samples' heterozygozity rate mean	60 individuals are removed
Relatedness	$\hat{p}$ of 0.125 (i.e., third degree relatives)	538 individuals are removed
QC Positive Individuals and SNPs		
4,452 individuals		
2,024 cases	2,428 controls	
Total number of SNPs: 305,250		

3.1.2 Generation of artificial genomes

Generative adversarial network (GAN) architecture details

In this study, the generative adversarial network (GAN) model architecture proposed by Yelmen et al. (2021) [5] was used, with the code accessible at `https://gitlab.inria.fr/ml_genetics/public/artificial_genomes`. The GAN architecture consisted of two fully connected networks: a generator and a discriminator. The generator included an input layer with the size of 600, two hidden layers with the size proportional to the number of SNPs as the rounded value of $SNP\_number/1.2$ and $SNP\_number/1.1$ correspondingly, and an output layer with the size equal to the number of SNPs. The discriminator included an input layer with the size equal to the number of SNPs, two hidden layers with the size proportional to the number of SNPs as the rounded value of $SNP\_number/2$ and $SNP\_number/3$ correspondingly, and an output layer with a single node. The activation function of the generator output layer is tanh, while the activation function of the discriminator output layer is sigmoid. Input and hidden layers' outputs have LeakyReLU activation function ($leaky\_alpha$ parameter = 0.01, $L2$ regularization parameter = 0.0001). [5] The network was implemented using python-3.6, Keras (2.4.4) deep learning library with TensorFlow backend [49], pandas (0.23.4) [50] and numpy (1.16.4) [51]. Please see Figure 3.1 for architecture details.

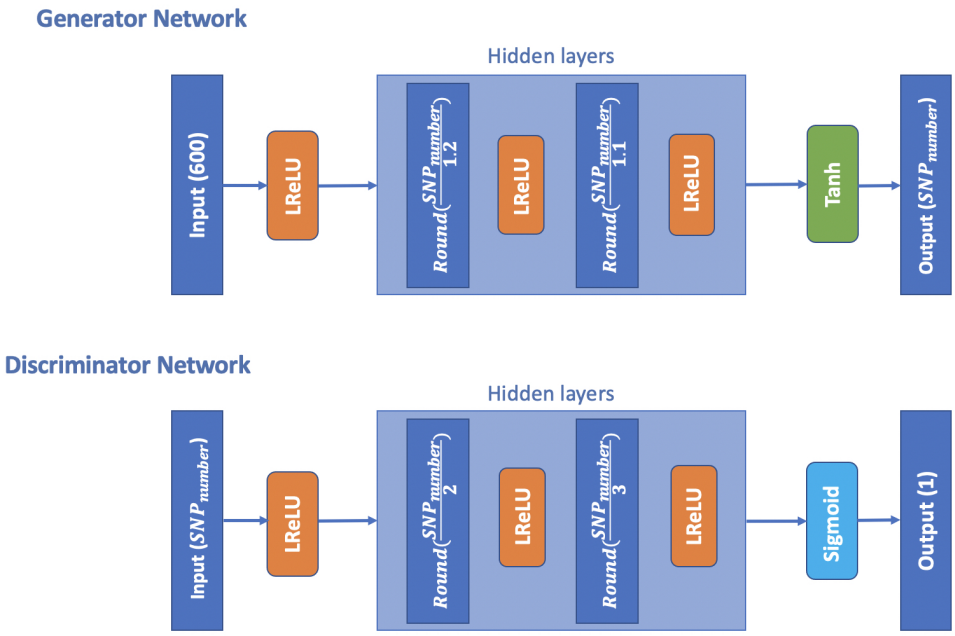

Figure 3.1: Generative adversarial network model's architecture used for artificial genomes generation

GAN training data

QC-positive individuals and SNPs from Estonian data were used as training data for the GAN model. The model training was conducted on cases and controls separately with the help of High Performance Computing Center of the University of Tartu. Data was represented in the following format: rows are haplotypes (instances) and columns are SNPs (features), with two rows representing haplotypes from one individual and each column representing 1 position. Alleles at each position are encoded in the binary format (with "0" and "1"). A latent vector of size 600 derived from a Gaussian distribution with zero mean and unit-variance served as an input for the generator. [5]

GAN training process

The discriminator and combined GAN training steps (gradient descent iterations) were chosen in a 1:1 ratio. For both the discriminator and the combined GAN, Adam optimization algorithm with binary cross-entropy loss function was used for training. The discriminator learning rate parameter was set to 0.00008 and the combined GAN learning rate was set to 0.00001. Training to test dataset ratio was 3:1, with a batch size of 32. During each batch of training, when only the discriminator is trained, smoothing to the real labels was applied to make the discriminator more generalized by adding numbers with random uniform distribution via $numpy.random.uniform$ within the 0.0 to 0.1 interval. Generated outputs were rounded to 0 or 1. [5]

During the training process, checkpoints were performed after every 200 epochs. At each checkpoint, models' parameters were saved, and AGs were produced in the form of either cases or controls, depending on the training data. The training process was assessed by principal component analysis (PCA) of AGs with real genomes, and the alignment of real and artificial data was

visually examined. The PCA analysis was done using scikit-learn (0.23.2) python library [52].

Initially, based on the training procedure of the previous study [5], we tried training in 10,000 SNPs long genomic chunks. However, the model convergence was not observed after 50,000 epochs. Attempts to train the model with 2,000 SNPs-length chunks failed to achieve the desired coherency after 50,000 epochs for some particular chunks as well. After cutting chromosomes into 1,000 SNPs, based on PCA results, AGs become visually coherent with real genomes at around 5,000 epochs, with some variations between chunks. We detected that at 10,000 epochs, all chunks seemed to be trained. Therefore, we proceeded with the model training based on 1,000 SNPs length and 10,000 epochs. All 22 chromosomes were trained using this approach. After the training process was completed, AG chunks with the most visually coherent PCA results were chosen and concatenated end to end (stitched) to form full chromosomes. It is important to note that AG chunks are produced by random sampling from the data distribution learned by the model, thus the model's output haplotype order does not match the training data. In other words, the selection of an artificial chunk for a given position in the chromosome was done randomly. Chopping of real genomes and stitching of AGs were done with customary python scripts (see Appendices A and B) using python-3.8 and pandas (1.1.4) [50].

Validation of generated AGs

After obtaining artificial chromosomes (chromosomes 7, 10, and 22) from the generated chunks, PCA analysis with real and AGs was done in order to visualize the similarity of stitched AGs with real ones. In addition, PCA analyses with cases and controls for real as well as AGs were performed for comparison of relationships between cases and controls.

3.1.3 Minor allele frequency correlation analysis

Minor allele frequency (MAF) analysis, stratified for cases and controls, was performed using PLINK (1.9) [48]. With the help of pingouin (0.3.11) python statistical library [53], the correlation of case MAF, as well as control MAF between real genomes and AGs, was estimated using the coefficient of determination (R^2). The correlation analysis between case and control MAF was also performed for artificial and real genomes separately. Results were visualized using seaborn (0.11.1) python visualization library [54] in the form of a joint plot.

3.1.4 Genome-wide association study

GWAS analyses of both real and AGs were conducted using PLINK version 1.9 [48]. For association analyses, 1 degree-of-freedom chi-square allelic test was used. Additionally, logistic regression analyses were conducted, allowing for the usage of covariates. The first 10 multi-dimensional scaling (MDS) components, as suggested by the tutorial [6], were computed from the data and used as covariates. The summary statistics in the Estonian T2D dataset were obtained from both association analyses and logistic regression with top 10 MDS components as well as sex and age as additional covariates. We separately performed both association analysis and logistic regression on real and artificial chromosomes 7 and 10, using only the top 10 MDS components as covariates for equal comparison as sex and age information is unavailable for artificial genomes. Obtained results were visualized in the form of Manhattan plots using QMplot python library available at `https://github.com/ShujiaHuang/qmplot`. To account for multiple testing, statistical significance genome-wide threshold was adjusted according to the Bounferroni correction ($0.05/n$ of SNPs tested). In our case, the genome-wide significance

threshold for GWAS conducted on all 22 chromosomes was equal to $0.05/305250 = 1.64e^{-7}$ (or $P < 6.79588 \times 10^{-8}$). Common 5×10^{-8} $P\text{-}value$ threshold was also displayed as a less stringent suggestive threshold. To prevent getting zero $P\text{-}values$ ($p\_value = 0$), which would make $-log_{10}$ conversion impossible, a $1e^{-99}$ ($P < 99 \times 10^{-8}$) minimal $P\text{-}value$ threshold was implemented. Additionally, quantile-quantile (Q-Q) plots showing the expected and obtained $-log_{10}(p\_value)$ distribution and computed genomic inflation (λ) factor were created using the mentioned QMplot library.

3.2 Results

3.2.1 Generation of artificial genomes

As mentioned in the "Materials and Methods" section 3.1, attempts to train the GAN model based on 10,000 and 2,000 SNP chunks were not successful (Figure 3.2). On the example of chromosome 16 (Figure 3.2, A), which has 8,684 positions, we can observe that artificial genomes are widely distributed throughout the plot even after 49,800 epochs, not covering the real genomes cluster. Splitting the data into 2,000 SNP chunks improved the results slightly, but in several cases, such as the first chunk of chromosome 22 (case) (Figure 3.2, B), artificial genomes were concentrated in one cluster, implying a weak coherency with real genomes.

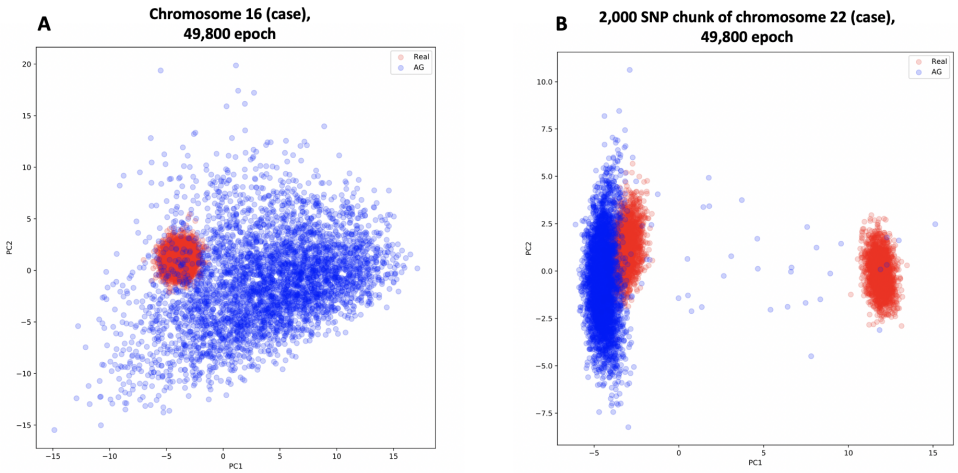

Figure 3.2: PCA visualization of the unsuccessful training sessions. The X-axis represents the first principal component (PC1), the Y-axis represents the second principal component (PC2). Real genomes are displayed in red color, while artificial genomes (AG) are displayed in blue color. (A) Chromosome 16 (case, whole chromosome) training at 49,800 epoch. (B) Chromosome 22 (case) first 2,000-length SNP chunk training at 49,800 epoch.

Training process based on 1,000 SNP long chunks resulted in the artificial genomes visually comparable to the real genomes before 5,000 epochs, indicating similar population structures and possibly a good fit (Figure 3.3). Therefore, this training approach was used to train all 22 chromosomes. For another example of chunk-based PCA results, please see Appendices C, D.

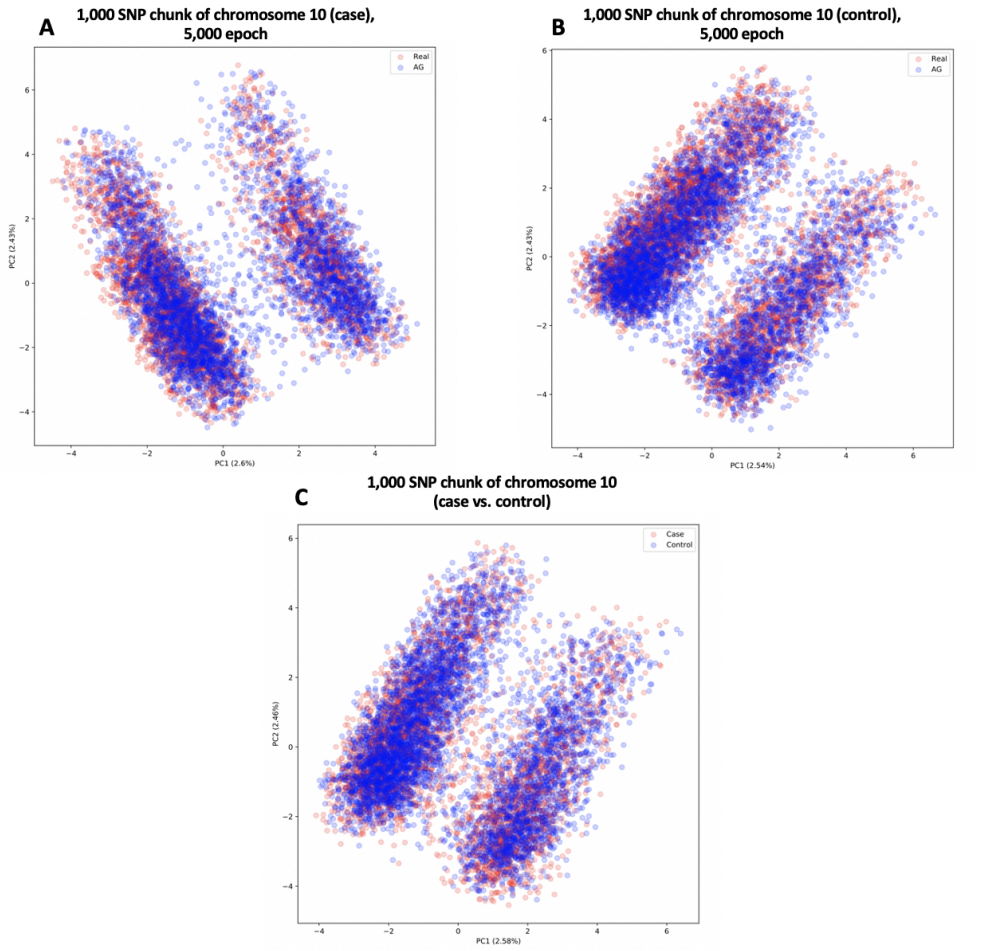

Figure 3.3: PCA analysis of 1,000 SNP chunks (chromosome 10, chunk 5). The X-axis represents the first principal component (PC1), the Y-axis represents the second principal component (PC2). For subplots A-B, real genomes are displayed in red color, while artificial genomes (AG) are displayed in blue color. For subplot C, case artificial genomes are displayed in red color, while control artificial genomes are displayed in blue color. (A) Chromosome 10 (case) fifth chunk training at 5,000 epoch. (B) Chromosome 10 (control) fifth chunk training at 5,000 epoch. (C) Chromosome 10 fifth chunk artificial genomes cases with controls.

After the training was completed, generated chunks were stitched back to whole chromosomes for subsequent analysis. To assess the overall quality of the produced artificial genomes, we performed the PCA analysis of stitched artificial genomes with real genomes (chromosomes 7, 10, 22). Here, we will present results for chromosome 10. Please see Appendices F, H for chromosome 7, and Appendices E, G for chromosome 22 PCA analysis results. Interestingly, the PCA analysis of the chromosomes demonstrated that artificial genomes clustered differently from the real genomes (Figure 3.4), although separate chunks were highly overlapping (Figure 3.3). It is important to note that the first principal component (PC1) displays the most variation within the data, while the second principal component (PC2) displays the second most variation. As a result, differences between clusters along the X-axis (PC1) are greater than differences along the Y-axis (PC2).

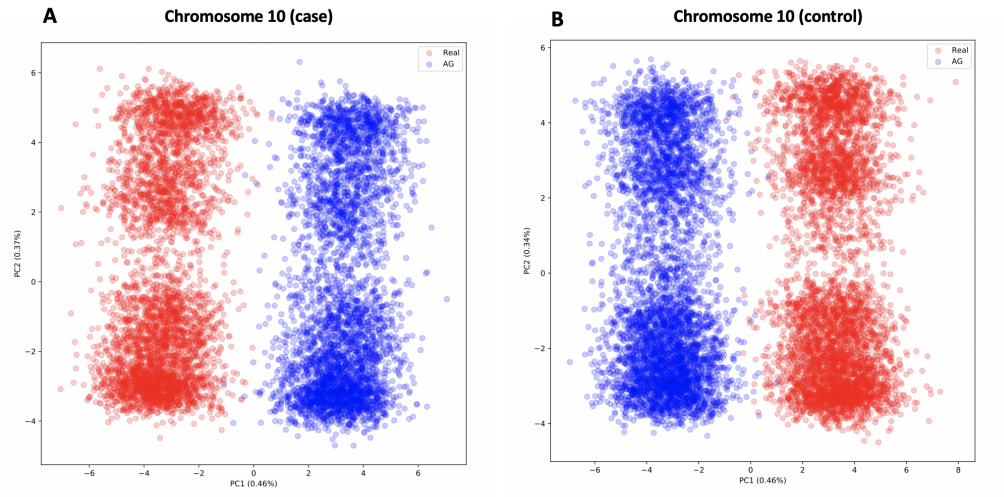

Figure 3.4: PCA analysis of real genomes with stitched artificial genomes (chromosome 10). The X-axis represents the first principal component (PC1), the Y-axis represents the second principal component (PC2). The percentage of variance explained by each principal component is displayed on the axis labels. Real genomes are displayed in red color, while artificial genomes (AG) are displayed in blue color. (A) Chromosome 10 (case). (B) Chromosome 10 (control).

To confirm that the difference between artificial and real genomes on the PCA plots is due to factors other than training pre- and postprocessing (chopping and random stitching), we replicated the same processes with real genomes. The PCA analysis of the real genomes with the same genomes which were cut into 1,000 SNP chunks, randomly shuffled, and stitched back, demonstrated that random stitching of shuffled chunks produces only some negligible differences (Figure 3.5).

Another PCA analysis was performed with cases and controls from real genomes, displaying expected overlapping results (Figure 3.6, A). However, PCA analysis with cases and controls from artificial genomes showed that case instances are located far from the controls (Figure 3.6, B).

By performing PCA every time a new chunk is introduced to a chromosome during the stitching process, we can see that the more chunks that are added, the greater the differences between the real and artificial genome clusters become (Figure 3.7). The same outcomes can be observed from the similar demonstration for differences between artificial genomes cases and controls: the more chunks are added, the larger differences we can see between case and control clusters (Appendix I). Excluding random chunks from stitched artificial genomes or chunks corresponding to chromosomal ends did not result in any positive improvements. This gives support to the idea of systematic error accumulation originating from single chunks.

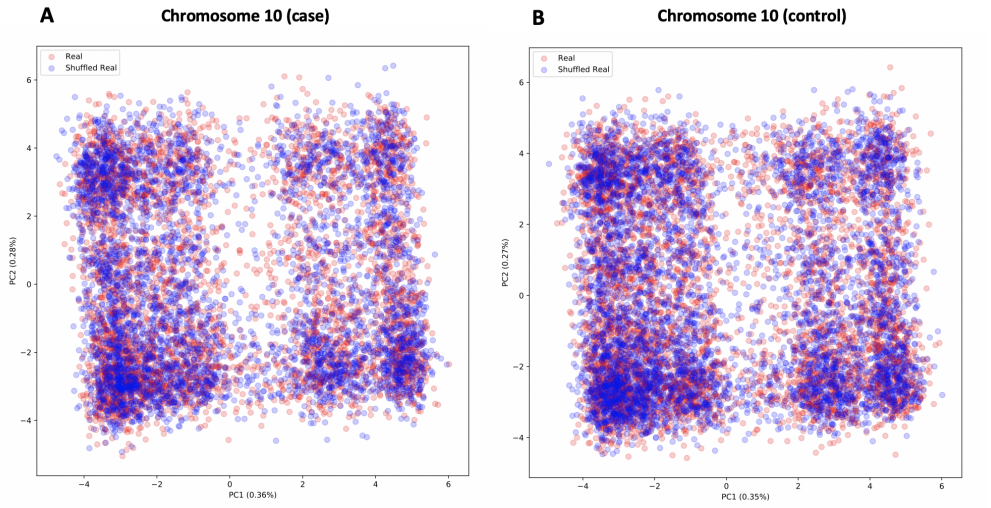

Figure 3.5: PCA analysis of real genomes with the same real genomes after the training pre- and post-processing (chromosome 10). The X-axis represents the first principal component (PC1), the Y-axis represents the second principal component (PC2). The percentage of variance explained by each principal component is displayed on the axis labels. Real genomes are displayed in red color, while shuffled real genomes are displayed in blue color. (A) Chromosome 10 (case). (B) Chromosome 10 (control).

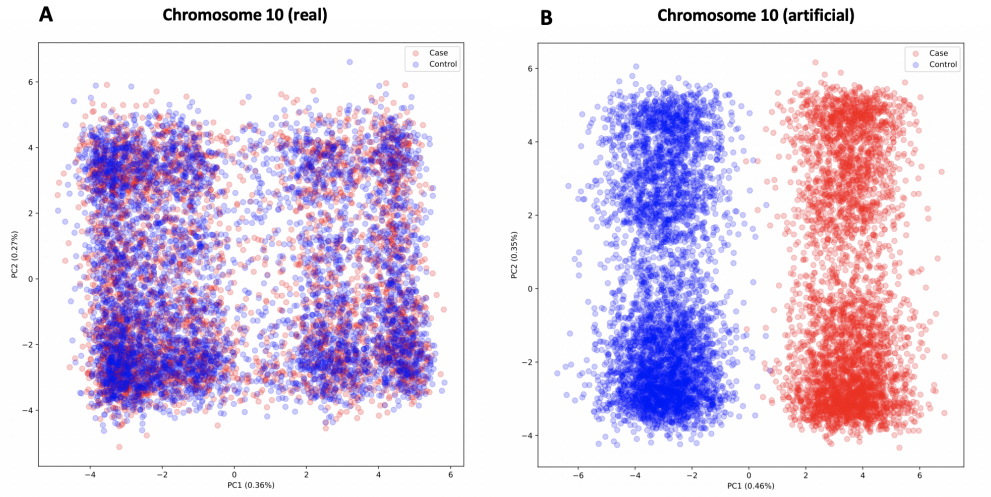

Figure 3.6: PCA analysis of cases with controls (chromosome 10). The X-axis represents the first principal component (PC1), the Y-axis represents the second principal component (PC2). The percentage of variance explained by each principal component is displayed on the axis labels. Cases are displayed in red color, while controls are displayed in blue color. (A) Chromosome 10 real genomes. (B) Chromosome 10 artificial genomes.

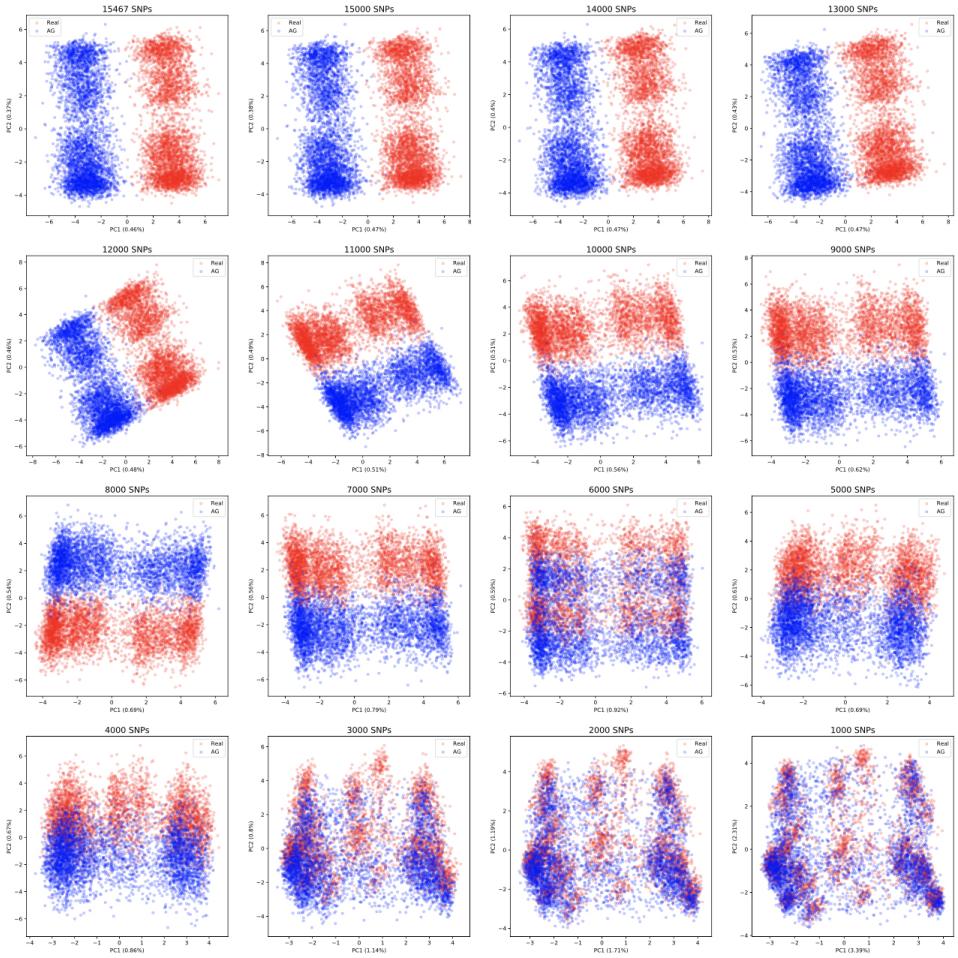

Figure 3.7: Demonstration of a systematic difference accumulation between real and artificial genomes during the stitching process. The analysis was performed on the chromosome 10 (case). Subplots represent PCA of different number of SNPs coming from different number of chunks stitched together. First subplot with 16,784 SNPs represents the full chromosome, while the last subplot with 1,000 SNPs represents one (first) training chunk. On each subplot, the X-axis represents the first principal component (PC1), the Y-axis represents the second principal component (PC2). The percentage of variance explained by each principal component is displayed on the axis labels. Real genomes are displayed in red color, while artificial genomes are displayed in blue color.

3.2.2 Minor allele frequency correlation analysis

Minor allele frequency (MAF) correlation analysis between the real and artificial genomes demonstrated that there is a positive correlation for both case MAF and control MAF (Figure 3.8). For the case MAF, the coefficient of determination (R2) is equal to 0.891107, indicating that approximately 89.1% of real genomes case MAF variance can be predicted from the artificial genomes case MAF (Figure 3.8, A). For the control MAF, the coefficient of determination (R2) is equal to 0.895914, indicating that approximately 89.6% of real genomes control MAF variance can be predicted from the artificial genomes control MAF (Figure 3.8, B).

An additional correlation analysis between the case MAF and control MAF conducted separately for the artificial and real genomes revealed that, although there is a positive correlation,

the MAF differences between cases and controls are not the same. For the real genomes (Figure 3.8, C), differences in MAFs between cases and controls are moderate, depicted by a narrow band in the relationship plot. For the artificial genomes (Figure 3.8, D), differences in MAFs between cases and controls are considerably higher than for the real genomes, depicted by a wider band in the relationship plot.

For the real genomes, the coefficient of determination (R2) is equal to 0.99524, indicating that approximately 99.5% of the case MAF variance can be predicted from the control MAF (Figure 3.8, C). For the artificial genomes, the coefficient of determination (R2) is lower and equal to 0.884831, indicating that approximately only 88.5% of the case MAF variance can be predicted from the control MAF (Figure 3.8, D).

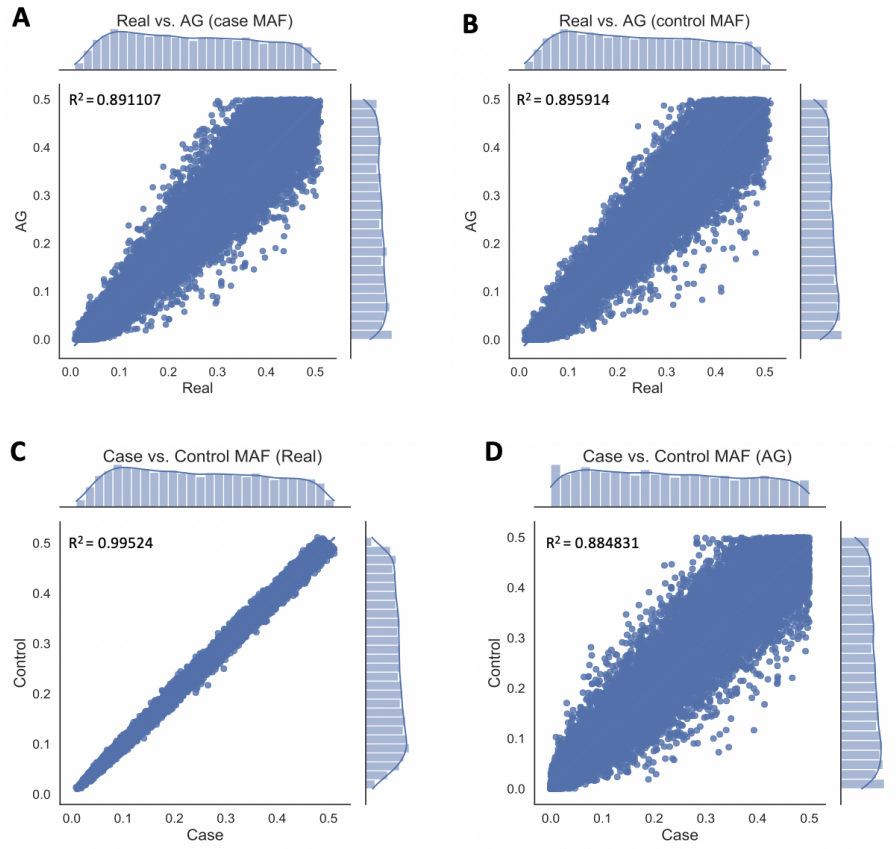

Figure 3.8: Minor allele frequency (MAF) correlation analysis between real and artificial genomes (chromosome 10). (A) Case MAF correlation between real and artificial genomes. (B) Control MAF correlation between real and artificial genomes. (C) Real genomes MAF correlation between cases and controls. (D) Artificial genomes (AG) MAF correlation between cases and controls. The X-axis represents MAF of real genomes (A, B) or cases (C, D), while the Y-axis represents MAF of artificial genomes (AG) (A, B) or controls (C,D). The marginal distributions are indicated as histograms on the sides. The coefficient of determination (R2) is also displayed on the plots.

We additionally decided to conduct the allele frequency correlation analysis for a single artificial genomes chunk, which was visually coherent with real genomes on the PCA plot. Similar results were obtained: despite the positive correlation, MAF differences between cases and

controls were higher for artificial genomes (Figure 3.9). For the case MAF, the coefficient of determination (R2) is equal to 0.946372, indicating that approximately 94.6% of real genomes case MAF variance can be predicted from the artificial genomes case MAF (Figure 3.9, A). For the control MAF, the coefficient of determination (R2) is equal to 0.943425, indicating that approximately 94.3% of real genomes control MAF variance can be predicted from the artificial genomes control MAF (Figure 3.9, B). For the real genomes, the coefficient of determination (R2) is equal to 0.997478, indicating that approximately 99.7% of the case MAF variance can be predicted from the control MAF (Figure 3.9, C). For the artificial genomes, the coefficient of determination (R2) is lower and equal to 0.940959, indicating that approximately only 94.1% of the case MAF variance can be predicted from the control MAF (Figure 3.9, D). For all 6 correlation analyses, the achieved power of the test is equal to 1, meaning that we can be confident in the obtained results and detected correlation. For other examples of MAF correlation analysis results please see Appendices J, K.

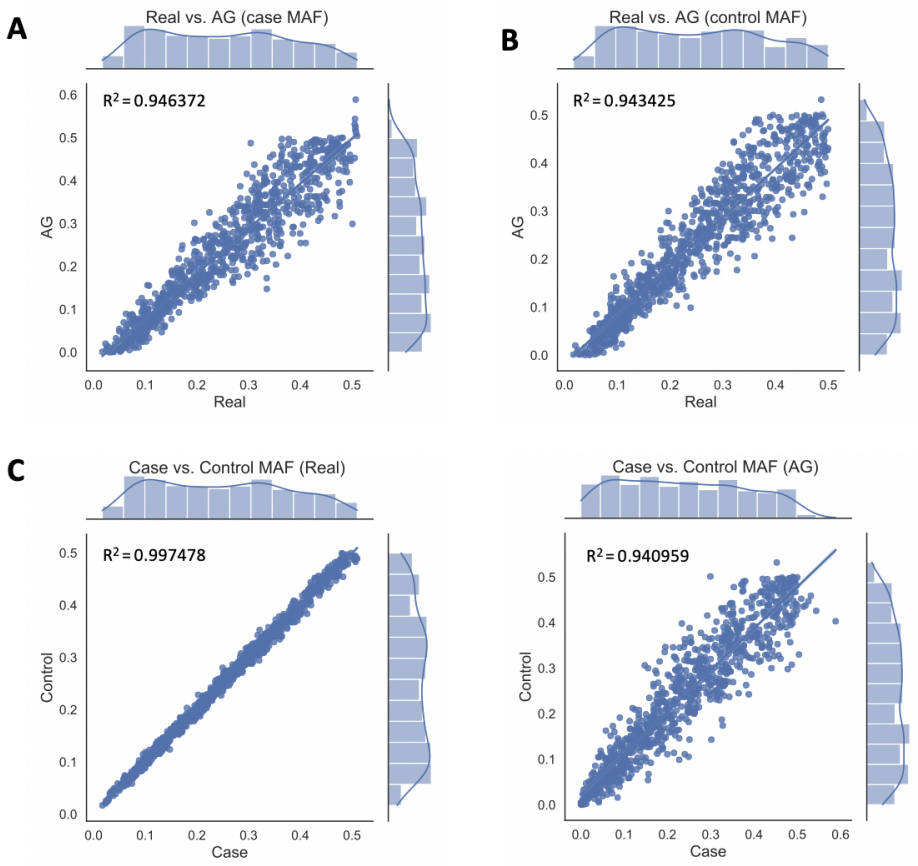

Figure 3.9: Minor allele frequency (MAF) correlation analysis between real and artificial genomes (chromosome 10, chunk 5). (A) Case MAF correlation between real and artificial genomes. (B) Control MAF correlation between real and artificial genomes. (C) Real genomes MAF correlation between cases and controls. (D) Artificial genomes (AG) MAF correlation between cases and controls. The X-axis represents MAF of real genomes (A, B) or cases (C, D), while the Y-axis represents MAF of artificial genomes (AG) (A, B) or controls (C,D). The marginal distributions are indicated as histograms on the sides. The coefficient of determination (R2) is also displayed on the plots.

3.2.3 Genome-wide association study

Estonian data GWAS

Association analysis of Estonian T2D data (Figure 3.10, A) resulted in 8 statistically significant hits that crossed a suggestive threshold: rs7903146, rs6459136, rs2055611, rs10500951, rs1554116, rs11652788, rs295869, rs12255372, and only 1 hit (rs7903146) crossed the more stringent genome-wide significance threshold. Two hits, rs7903146 and rs12255372, were previously reported in the literature [55, 56] and are stated in the ClinVar NCBI database [57] as associated with type 2 diabetes (rs7903146 accession: SCV000028043.4; rs12255372 accession: SCV000028044.3). The calculated genomic inflation factor (λ) equals 1.056, which is acceptable. Logistic regression with gender, age, and top 10 MDS components as covariates (Figure 3.10, B) displayed 5 hits crossed the suggestive genome-wide threshold: rs288864, rs7903146, rs2158091, rs6580921, rs11652788. Only 1 SNP (rs7903146) was previously reported in the ClinVar NCBI database [57] as associated with T2D. No SNPs crossed a more stringent genome-wide significant threshold. λ value equals 1.027, which is acceptable and even closer to 1, meaning that multidimensional scaling (MDS) components were helpful to account for underlying population structure. Quantile-quantile plots depict that the distribution of the observed p-values is overall coherent with the expected distribution. Most SNPs are under the null hypothesis, while the small fraction deviates at smaller p-values, indicating statistically significant candidates.

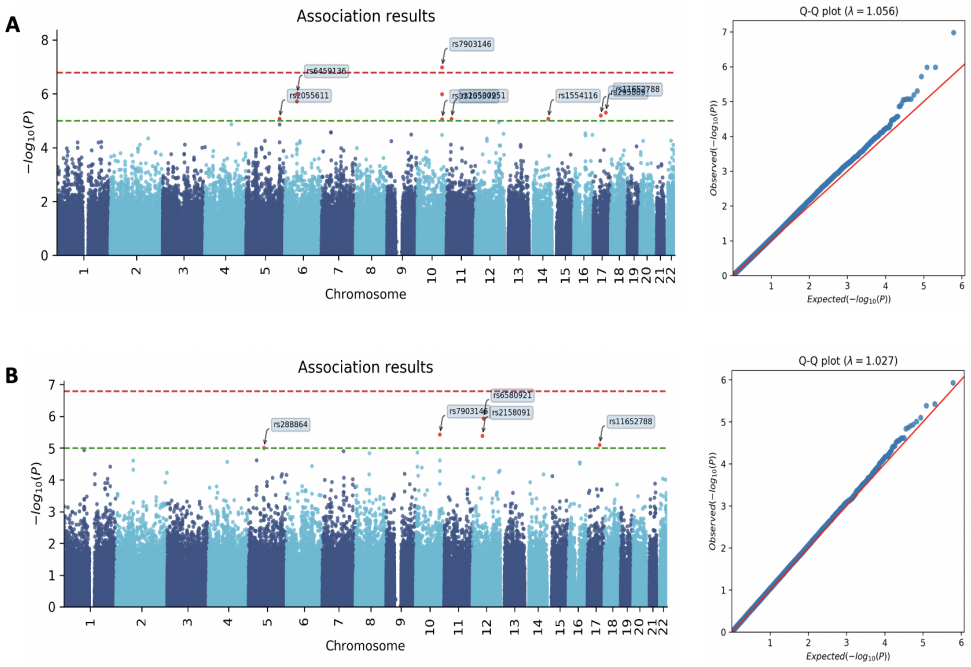

Figure 3.10: GWAS analysis of Estonian type 2 diabetes (T2D) data. Results are represented in the form of manhattan plot. The X-axis on the figure indicates haplotypes from each tested region of the genome, which are organized by chromosomes shown in different color blocks. The Y-axis indicates p-values in the scale of negative common logarithm. Green line represents standard suggestive threshold ($P<5 \times 10^{-8}$), while red line represents more stringent, adjusted for Bonferroni correction, genome-wide threshold ($P<6.79588 \times 10^{-8}$), creating a statistical significance borderline. Top hits crossing the suggestive threshold are marked on the plot. Results are supplemented with quantile-quantile (Q-Q) plots, where genomic inflation factor (lambda) is displayed in the title. (A) Association analysis. (B) Logistic regression analysis with age, sex and top 10 multidimensional scaling (MDS) components as covariates.

Artificial genomes GWAS

Despite the fact that artificial genomes were shown to differ from real genomes in PCA and allele frequency correlation analyses, we decided to proceed with the GWAS and investigate the behavior of artificial genomes further and performed GWAS analysis for chromosome 10 (Figure 3.11). Association analysis (Figure 3.11, A) resulted in almost all positions being considered as highly statistically significant. The calculated genomic inflation factor (λ) equals 25.975, which is not acceptable. Quantile-quantile (Q-Q) plot depicts the presence of extremely small p-values, indicating that the results are highly inflated (many false positives). Logistic regression with top 10 MDS components as covariates (Figure 3.11, B) resulted in almost all positions being considered as statistically not significant. The calculated genomic inflation factor (λ) equals 14.241, which is not acceptable. Quantile-quantile plot depicts the presence of extremely high p-values, indicating that the results are highly deflated (many false negatives). The analysis was repeated for chromosome 7, following the same results patterns (see Appendix L).

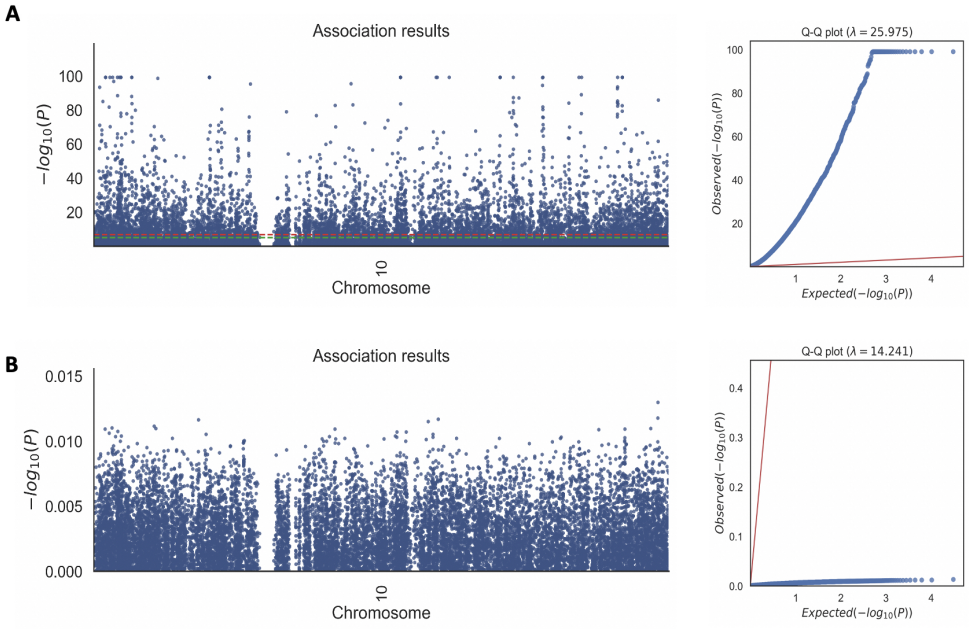

Figure 3.11: GWAS analysis of artificial genomes chromosome 10 data. Results are represented in the form of manhattan plot. The X-axis on the figure indicates haplotypes from each tested region of the chromosome 10. The Y-axis indicates p-values in the scale of negative common logarithm. Green line represents standard suggestive threshold ($P<5 \times 10^{-8}$), while red line represents more stringent, adjusted for Bonferroni correction, genome-wide threshold ($P<6.79588 \times 10^{-8}$), creating a statistical significance borderline. Top hits crossing the suggestive threshold are marked on the plot. Results are supplemented with quantile-quantile (Q-Q) plots, where genomic inflation factor (λ) is displayed in the title. (A) Association analysis. (B) Logistic regression analysis with top 10 multidimensional scaling (MDS) components as covariates.

Results of GWAS analysis on a single chunk in chromosome 10 were still highly inflated (Figure 3.12). After the addition of the top 10 MDS components as covariates, the significance decreased, but results did not become deflated (Figure 3.12, B), as in the case with stitched artificial genomes (Figure 3.11, B). The genomic inflation factor (lambda) for association analysis equals 23.222 and for logistic regression equals 9.222, which is not acceptable. For both association analysis and logistic regression, the quantile-quantile plot depicts the presence of extremely small p-values (many false positives). The analysis was repeated for chunk 13 of chromosome 7, following the same results patterns (see Appendix M).

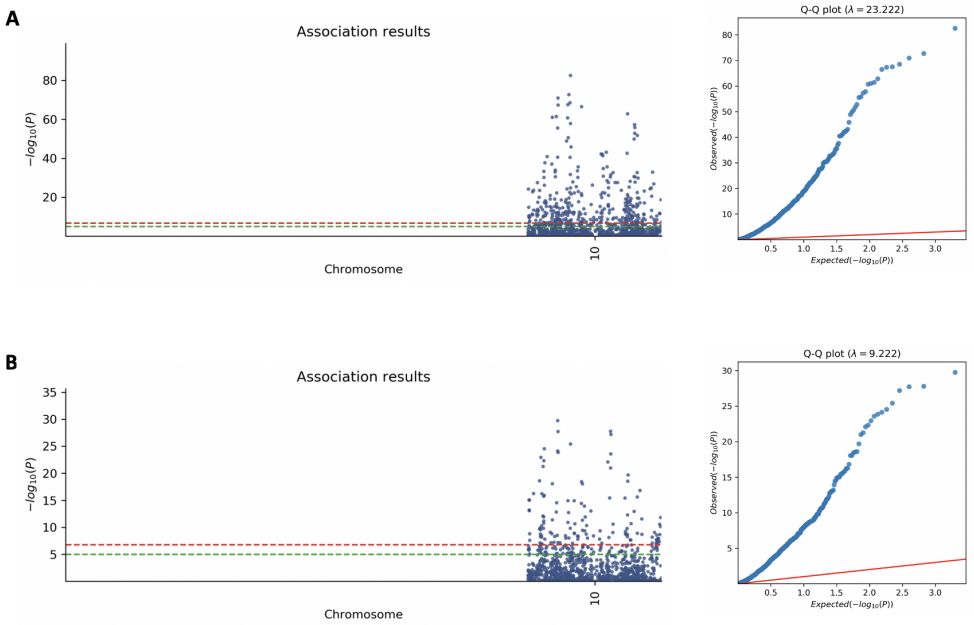

Figure 3.12: GWAS analysis of artificial genomes chromosome 10, chunk 5 data. Results are represented in the form of manhattan plot. The X-axis on the figure indicates haplotypes from each tested region of the chunk 5. The Y-axis indicates p-values in the scale of negative common logarithm. Green line represents standard suggestive threshold ($P<5 \times 10^{-8}$), while red line represents more stringent, adjusted for Bonferroni correction, genome-wide threshold ($P<6.79588 \times 10^{-8}$), creating a statistical significance borderline. Top hits crossing the suggestive threshold are marked on the plot. Results are supplemented with quantile-quantile (QQ) plots, where genomic inflation factor (λ) is displayed in the title. (A) Association analysis. (B) Logistic regression analysis with top 10 multidimensional scaling (MDS) components as covariates.

3.3 Discussion

Our experimental work started with testing several training approaches, mostly with varying genomic chunk lengths. Despite the fact that in the previous study by Yelmen et al. (2021) training with 10,000 SNPs was successful, on Estonian T2D data, models could not converge (Figure 3.2) and required a substantially smaller number of variants to capture the data distribution (Figure 3.3). In Yelmen et al. (2021), the training was done with the genomic region of chromosome 15 from 27379578 to 29625035, which had 10,000 SNPs. The exact same region in our SNP array data has 170 SNPs. In addition, the data in Yelmen et al. (2021) contained many fixed alleles (a phenomenon when only one allele exists for the particular locus). This suggests that Estonian T2D data has higher informational content, making it more complex for the model to learn.

Eventually, our training approach required splitting chromosomes into 1,000 SNPs chunks. For every single chunk, artificial genomes were largely overlapping with the real ones, and artificial cases and controls also showed similar population structures (Figure 3.3, Appendices C, D). Since artificial genomes are produced by random sampling from the data distribution learned by the model, the model's output haplotype order does not match the training data. Therefore, during the stitching, haplotypes are also concatenated randomly. This could lead to the disrup-

tion of linkage disequilibrium (LD) blocks. In Yelmen et al. (2021) paper, the authors suggest conducting training on "approximately independent LD blocks" to reduce the possible disruption [58]. However, since genetic association tests are mostly based on the differences in allele frequencies between cases and controls [59], we thought that possible LD blocks disruption should not affect the final results of our study. Despite this, we assumed that the disruption of LD blocks spanning multiple regions might be the reason for large differences between real and stitched artificial genomes observed on the PCA (Figure 3.4). To test this possibility, we first chopped real chromosomes into 1000 SNP chunks and concatenated them randomly to produce similarly processed real genomes. PCA demonstrated visually coherent results with processed and not processed real genomes (Figure 3.5).

We then assumed that the problem might be caused due to the complexity of chunks corresponding to the chromosomal ends, or by some specific problematic chunks. To verify it, we ran a PCA analysis after excluding a few first and last chunks, as well as one chunk at a time in random order. However, in both cases, artificial genomes were still clustering distinctly, eliminating this hypothesis. Assuming that there might be small variations within single chunks that are not detected by the PCA, we analyzed gradually stitched artificial genomes. The PCA demonstrated the presence of a systematic difference between real and artificial genomes, which is amplified during the stitching process (Figure 3.7). Interestingly, additional PCA analysis between cases and controls revealed a similar effect: a systematic error was accumulating as multiple regions were concatenated, resulting in significant differences (which can not be detected visually with PCA for 1,000 SNP chunks) between artificial cases and controls in the end (Appendix I).

Another interesting detail we noticed was the presence of distinct clusters on the PCA plots of real genomes, not connected with case and control phenotypes (Figure 3.3). Since our data were obtained using different genotyping platforms, such clusters could appear due to the so-called "batch effect", which causes non-biological differences and variation within the data [60], although the exact reason for this observed genomic structure remains to be investigated in the future.

Even though minor allele frequency analysis showed a positive correlation between the real and artificial genomes (Figure 3.8, A, B), the allele frequency difference between cases and controls was larger for the artificial genomes compared to real genomes (Figure 3.8, C, D). Allele frequency analysis on the single chunk revealed the same pattern (Figure 3.9). It is important to note that in the previous study by Yelmen et al. (2021), artificial genomes also demonstrated a positive correlation in allele frequencies with some deviations from the real genomes. While differences in allele frequencies can be beneficial in terms of diversity and privacy preservation, they can cause problems for the GWAS as small differences in allele frequencies between cases and controls can easily result in statistical significance with a large enough sample size [59]. We can assume that these differences between cases and controls are the major issue that caused highly inflated GWAS results (Figure 3.11, A).

The differences in the PCA plots between cases and controls (Figure 3.6) also contribute to this explanation. GANs are non-deterministic models, which means that training outputs could be produced with some random variations, and since the training was done separately for cases and controls, this might also result in such differences. It is a well-known fact that population stratification (PS), which is the presence of several sub-populations in the data, can confound association studies and lead to false-positive correlations or mask true associations due to the

differences in allele frequencies between sub-populations [6]. Furthermore, although the significance of choosing cases and controls from the same breeding populations has been known for decades, recent large-scale GWAS highlighted that when cases and controls have different substructures, the number of false-positive associations is inflated [59].

In our situation, the first principal component's variance in Figure 5B equals 0.46%, which suggests the presence of population structure. Therefore, we can consider case and control groups as two fairly distinct discrete populations. We could imagine a hypothetical example of conducting a GWAS with Estonian and African populations. In this case, the analysis would reveal many spurious associations. That is why it is important to account for population structure and include corresponding covariates. Particularly, adjusted logistic regression is considered to be one of the best solutions to account for population stratification [61]. After the addition of the top 10 MDS components, we observed that statistical significance drastically decreased, and results even became deflated (Figure 3.11, B). The genomic inflation factor (λ) was enormously high and Q-Q plots demonstrated a strong deviation from the uniform distribution, indicating the excess of false positives (or negatives, when accounted for population structure). One possible explanation for this behavior is that statistical significance detected for multiple positions in artificial genomes is strongly correlated with the underlying pseudo-population structure. Besides population stratification, a large lambda can be caused by other types of systematic errors between the case and control groups, such as nonrandom differences in DNA quality between study groups, genotyping errors, or other unmeasured confounding factors [14].

Interestingly, GWAS analysis results on the single chunk were also highly inflated (Figure 3.12, A), despite the demonstration of a highly similar structure with real genomes and overlapping of artificial cases with controls on PCA plots (Figure 3.3). Moreover, after the addition of the top 10 MDS components as covariates, logistic regression results were less statistically significant but did not become deflated (Figure 3.12, B). Since minor allele frequency analysis still demonstrated larger differences between cases and controls (Figure 3.9, C, D), we assumed that PCA could not capture these subtle differences for a single chunk, but their presence can be observed when chunks are stitched (Appendix I).

Finally, GWAS on Estonian T2D data revealed several statistically significant SNPs that were not previously reported. It is important to note that SNPs found to be statistically correlated with a disease are not always causal variants, and further analysis is needed to determine their biological significance. What is more, we did not include body-mass index (BMI) as a covariate in our study, which is crucial for T2D GWAS. As a result, we are unable to draw any conclusions about the relationship of these SNPs to T2D, and we will leave these variants for further examination to other researchers.

More into future research directions and possible improvements, a first step will be to explore alternative training strategies. First and foremost, since differences between cases and controls should be relatively small, artificial genomes could be created with the model trained on cases and controls together. In the form of an additional column, binary class labels (for example, "0" for cases and "1" for controls) can be assigned to each haplotype in the training data. A similar approach was used by Yelmen et al. (2021) in order to check the ability of generative models to capture genotype-phenotype associations on the example of blue and brown eye colors. Hopefully, this training approach would help the model to minimize biologically irrelevant differences between cases and controls, while learning the important for the GWAS features of the two groups and better capturing the signal-to-noise ratio. Besides, splitting chromosomes

into even smaller chunks could assist the model in learning the underlying data structure and catching allele frequencies more precisely, avoiding the accumulation of systematic error when chunks are stitched, at the expense of the integrity of the haplotypic structure.

LD blocks preservation, on the other hand, could be important for the haplotype-based GWAS, which gains its popularity among the research community due to its superior ability in controlling false positives and detecting genotyped causal variants without relying on LD [62]. Tackling the issue of LD blocks disruption, training could be conducted on the genomes with some overlapping regions rather than simply breaking chromosomes into separate chunks. Eventually, produced artificial genomes could be stitched based on the overlapping windows, preserving the correlation between the variants and maintaining a correct genomic structure.

We could also take into account the possibility of varying other model parameters to make the training process more efficient. As discussed previously, mode collapse is a well-known problem in GANs. In our study, we could sometimes observe in certain instances that artificial genomes preferred to accumulate on one particular cluster, potentially creating more differences between the artificial cases and controls (for example, in Figure 3.3, we can observe more artificial genomes represented by a saturated blue color on the left of the real genomes cluster). There are several approaches to address this issue. For instance, if data is divided into n number of classes (in our case, n would be equal to 2, representing two classes: cases and controls), the discriminator could be trained to classify data samples into n + 1 classes (continuing with our study example, discriminator would classify the data instances into 3 classes, 2 of which represent classes in the data, cases, and controls, and an additional class corresponds to the data samples produced by the generator). Providing the labels of classes to the model. It was shown by Salimans et al. (2016) that such an approach could result in a better quality of produced samples. We could also try to make the discriminator model more complex. Since the loss function of the generator is more useful when the discriminator is powerful enough to distinguish between real and fake data, the generator would have to compete with a more "idealistic" discriminator, avoiding failing into one data mode.

Moreover, we could investigate the idea of applying transfer learning to our problem. With this approach, the model could be trained initially on some subset of chromosomal chunks, learning more general parameters, and then applied on another chunk portion, "fine-tuning" the parameters for learning more specific features. Although transfer learning in GANs is an ongoing research topic, it could decrease the training time and improve training stability by helping the model converge faster [63].

Finally, the evaluation of model performance and quality of produced artificial genomes could be more comprehensive. As our study shows, PCA might not be the best training criterion for the quality of produced artificial genomes, especially for some particular applications such as GWAS. As a possible solution, we could broaden the training assessment to several additional metrics. The model performance could be quantified by adopting a method similar to the one implemented by Wang et al. (2020), which uses the discriminator's accuracy as an evaluation metric. Moreover, additional analyses such as allele frequencies or LD correlations can be computed from the generated data and compared to the training data. Furthermore, although we conducted the training on QC-positive individuals and SNPs, generated artificial genomes might additionally be examined with some of the QC thresholds.

Another technically different approach would be to implement some derivative models of GANs

with certain improvements and additional features. For example, in the recent study by Repecka et al. (2021), researchers investigated the ability of GAN to create functional protein sequences. For this task, a special GAN architecture was developed specifically to learn patterns in long biological sequence data. It included convolutional filters with dilation, which increased the receptive field of the model and improved its ability to capture long-distance relationships, and layers with self-attention mechanism, which highlighted functionally important areas across the entire sequence. A similar model's structure on raw genotype matrix data probably could allow the model to train on the larger sequences while capturing important local sequence features such as LD blocks. Implementation of another GAN derivative model, Conditional GAN (CGAN) [64], through adding a conditional variable, could enable us to guide the data generation process, producing genomes with specific properties such as phenotypes. One important issue with artificial genomes is that they lack covariates, which are needed for the trustworthy GWAS analysis. Using a conditional GAN approach, it might be possible to create artificial genomes with, for example, BMI or specific age, given that input data were labeled. The same approach can be used to produce artificial genomes that belong to different sub-populations, which is often the case for the GWAS data. Moreover, artificial genomes with known properties could become a valuable asset in other applications. For instance, supervised machine learning applications, including ones focused on the post-GWAS analysis improvements and locus prioritization, would highly benefit from the availability of labeled simulated genomic data [18].

It is important to note that, besides the known limitations of artificial genomes and some points requiring further research ("known unknown"), there always exists "unknown unknown", some issues that we could not even take into account due to the complex nature of genomic data. Even though artificial genomes look similar to real genomes according to some set of characteristics, there is a possibility that they do not exhibit certain properties for various genomic applications.

Additionally, despite the overall promising applicability, artificial genomes as a solution for genomic data inaccessibility could be an example of "cracking a nut with a sledgehammer" in the case of GWAS. The problem of data availability is undeniably a limiting factor for new variant discovery and therefore must be addressed, but there may be other, easier strategies. In principle, allele frequencies from the genomic databases could be publicly released without any privacy concerns. Moreover, GWAS summary statistics, which usually contain effect size (odds ratio/beta), p-values, standard error, and MAF [6], are often accessible for the researchers [65]. A recent study by Yang et al. (2021) proposes a novel framework that can reconstruct allelic frequencies or genotypic counts of each SNP from the case-control GWAS summary statistics. This study serves as a good example of possible alternatives to artificial genomes.

On the other hand, providing sequence data enables scientists to do more than just statistical association analysis. Moreover, improved GWAS methods involving haplotypic information [62] can only be performed with sequence data, hence making AGs future candidates for such applications. As an outcome, if artificial genomes are proven to be applicable in GWAS, they should be tested further on a broader spectrum of variant discovery analysis applications.

One of the most critical factors for effective GWAS research is a genomic dataset with sufficient sample size. Nevertheless, sharing of genomic and phenotypic data is strictly controlled to protect human confidentiality, making it difficult for researchers to obtain high-quality datasets. In this work, the first steps were undertaken to investigate the use of AGs in GWAS as a potential alternative for inaccessible genomic data. Despite the fact that AGs did not perform well in GWAS, most likely due to differences in allele frequencies between case and control

groups, it is too early to withdraw them from the prospective GWAS applications. By training GAN separately for cases and controls, we investigated probably the most intuitive training option. However, as we found out, this approach may not be the best, especially considering the non-deterministic nature of GAN. As a result, further research, particularly focusing on other training approaches, should be pursued before drawing any firm conclusions. Hopefully, if proven to work, AGs may be able to assist researchers in conducting more and better GWAS by ”democratizing” genomic data, eventually contributing to complex disease research.

Summary

In this work, we tested the applicability of artificial genomes (AGs) in genome-wide association studies (GWAS) by analyzing AGs generated based on Estonian type 2 diabetes (T2D) data. The generative adversarial network (GAN) model was trained separately for two groups: disease cases and healthy controls. We found that, depending on the informational content of the data, the number of single nucleotide polymorphisms (SNPs) for the successful model training can vary. We tried several training approaches with different chunk lengths and eventually were able to generate 1,000 SNP-long artificial case and control genomic chunks. We showed that, for these SNPs chunks, artificial genomes were highly overlapping with real genomes on the PCA. Moreover, AGs case and control groups represented a similar population structure.

However, after stitching back to full chromosomes, we discovered that AGs represented a highly distinct population structure compared to real genomes. In addition, artificial cases and controls were shown to cluster differently. We assumed that the reason for such behavior could be a systematic difference between real and artificial genomes as well as cases and controls, which is amplified during the stitching process. Subsequently, we conducted minor allele frequency (MAF) correlation analysis, which revealed that differences in allele frequencies between cases and controls are larger for the AGs.

Nevertheless, we performed GWAS analysis both on the Estonian T2D data and AGs data. GWAS on Estonian data produced robust results, with two SNPs reported previously as associated with T2D. In contrast, AGs demonstrated poor performance in GWAS, as we expected due to the allele frequency difference between artificial case and control groups. We concluded that highly inflated GWAS results for the AGs are possibly due to large differences in allele frequencies between cases and controls and the observed pseudo-population structure. In addition, other training strategies, possible technical improvements, and future research directions were proposed to investigate the potential of AGs in GWAS further.

Bibliography

[1] Jennie Lin and Katalin Susztak. Complexities of understanding function from CKD-associated DNA variants. *Clinical Journal of the American Society of Nephrology*, 15(7), 2020.

[2] Vivian Tam, Nikunj Patel, Michelle Turcotte, Yohan Bossé, Guillaume Paré, and David Meyre. Benefits and limitations of genome-wide association studies, 2019.

[3] Jason Flannick and Jose C. Florez. Type 2 diabetes: Genetic data sharing to advance complex disease research, 2016.

[4] G. M. Harshvardhan, Mahendra Kumar Gourisaria, Manjusha Pandey, and Siddharth Swarup Rautaray. A comprehensive survey and analysis of generative models in machine learning. *Computer Science Review*, 38, 2020.

[5] Burak Yelmen, Aurélien Decelle, Linda Ongaro, Davide Marnetto, Corentin Tallec, Francesco Montinaro, Cyril Furtlehner, Luca Pagani, and Flora Jay. Creating artificial human genomes using generative neural networks. *PLOS Genetics*, 17(2):e1009303, 2 2021.

[6] Andries T. Marees, Hilde de Kluiver, Sven Stringer, Florence Vorspan, Emmanuel Curis, Cynthia Marie-Claire, and Eske M. Derks. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *International Journal of Methods in Psychiatric Research*, 27(2), 6 2018.

[7] Peter M. Visscher, Naomi R. Wray, Qian Zhang, Pamela Sklar, Mark I. McCarthy, Matthew A. Brown, and Jian Yang. 10 Years of GWAS Discovery: Biology, Function, and Translation, 2017.

[8] Melinda C. Mills and Charles Rahal. A scientometric review of genome-wide association studies, 2019.

[9] Lianchao Jin, Fuxiao Tan, and Shengming Jiang. Generative Adversarial Network Technologies and Applications in Computer Vision, 2020.

[10] Daniel R. Schrider and Andrew D. Kern. Supervised Machine Learning for Population Genetics: A New Paradigm, 4 2018.

[11] Jennie Lin and Kiran Musunuru. From Genotype to Phenotype: A Primer on the Functional Follow-up of Genome-Wide Association Studies in Cardiovascular Disease. *Circulation. Genomic and precision medicine*, 11(2), 2018.

[12] Anubha Mahajan, Min Jin Go, Weihua Zhang, Jennifer E. Below, Kyle J. Gaulton, Teresa Ferreira, Momoko Horikoshi, Andrew D. Johnson, Maggie C.Y. Ng, Inga Prokopenko, Danish Saleheen, Xu Wang, Eleftheria Zeggini, Goncalo R. Abecasis, Linda S. Adair, Peter Almgren, Mustafa Atalay, Tin Aung, Damiano Baldassarre, Beverley Balkau, Yuqian Bao, Anthony H. Barnett, Ines Barroso, Abdul Basit, Latonya F. Been, John Beilby, Graeme I. Bell, Rafn Benediktsson, Richard N. Bergman, Bernhard OBoehm, Eric Boerwinkle, Lori L. Bonnycastle, Noël Burtt, Qiuyin Cai, Harry Campbell, Jason Carey, Stephane Cauchi, Mark Caulfield, Juliana C.N. Chan, Li Ching Chang, Tien Jyun Chang, Yi Cheng Chang, Guillaume Charpentier, Chien Hsiun Chen, Han Chen, Yuan Tsong Chen, Kee Seng Chia, Manickam Chidambaram, Peter S. Chines, Nam H. Cho, Young Min Cho, Lee Ming Chuang, Francis S. Collins, Marilyn C. Cornelis, David J. Couper, Andrew T. Crenshaw, Rob M. Van Dam, John Danesh, Debashish Das, Ulf De Faire, George Dedoussis, Panos Deloukas, Antigone S. Dimas, Christian Dina, Alex S.F. Doney, Peter J. Donnelly, Mozhgan Dorkhan, Cornelia Van Duijn, Josée Dupuis, Sarah Edkins, Paul Elliott, Valur Emilsson, Raimund Erbel, Johan G. Eriksson, Jorge Escobedo, Tonu Esko, Elodie Eury, Jose C. Florez, Pierre Fontanillas, Nita G. Forouhi, Tom Forsen, Caroline Fox, Ross M. Fraser, Timothy M. Frayling, Philippe Froguel, Philippe Frossard, Yutang Gao, Karl Gertow, Christian Gieger, Bruna Gigante, Harald Grallert, George B. Grant, Leif C. Groop, Christopher J. Groves, Elin Grundberg, Candace Guiducci, Anders Hamsten, Bok Ghee Han, Kazuo Hara, Neelam Hassanali, Andrew T. Hattersley, Caroline Hayward, Asa K. Hedman, Christian Herder, Albert Hofman, Oddgeir L. Holmen, Kees Hovingh, Astradur B. Hreidarsson, Cheng Hu, Frank B. Hu, Jennie Hui, Steve E. Humphries, Sarah E. Hunt, David J. Hunter, Kristian Hveem, Zafar I. Hydrie, Hiroshi Ikegami, Thomas Illig, Erik Ingelsson, Muhammed Islam, Bo Isomaa, Anne U. Jackson, Tazeen Jafar, Alan James, Weiping Jia, Karl Heinz Jöckel, Anna Jonsson, Jeremy B.M. Jowett, Takashi Kadowaki, Hyun Min Kang, Stavroula Kanoni, Wen Hong L. Kao, Sekar Kathiresan, Norihiro Kato, Prasad Katulanda, Sirkka M. Keinanen-Kiukaanniemi, Ann M. Kelly, Hassan Khan, Kay Tee Khaw, Chiea Chuen Khor, Hyung Lae Kim, Sangsoo Kim, Young Jin Kim, Leena Kinnunen, Norman Klopp, Augustine Kong, Eeva Korpi-Hyövälti, Sudhir Kowlessur, Peter Kraft, Jasmina Kravic, Malene M. Kristensen, S. Krithika, Ashish Kumar, Jesus Kumate, Johanna Kuusisto, Soo Heon Kwak, Markku Laakso, Vasiliki Lagou, Timo A. Lakka, Claudia Langenberg, Cordelia Langford, Robert Lawrence, Karin Leander, Jen Mai Lee, Nanette R. Lee, Man Li, Xinzhong Li, Yun Li, Junbin Liang, Samuel Liju, Wei Yen Lim, Lars Lind, Cecilia M. Lindgren, Eero Lindholm, Ching Ti Liu, Jian Jun Liu, Stéphane Lobbens, Jirong Long, Ruth J.F. Loos, Wei Lu, Jianan Luan, Valeriya Lyssenko, Ronald C. WMa, Shiro Maeda, Reedik Mägi, Satu Männistö, David R. Matthews, James B. Meigs, Olle Melander, Andres Metspalu, Julia Meyer, Ghazala Mirza, Evelin Mihailov, Susanne Moebus, Viswanathan Mohan, Karen L. Mohlke, Andrew D. Morris, Thomas WMühleisen, Martina Müller-Nurasyid, Bill Musk, Jiro Nakamura, Eitaro Nakashima, Pau Navarro, Peng Keat Ng, Alexandra C. Nica, Peter M. Nilsson, Inger Njølstad, Markus M. Nöthen, Keizo Ohnaka, Twee Hee Ong, Katharine R. Owen, Colin N.A. Palmer, James S. Pankow, Kyong Soo Park, Melissa Parkin, Sonali Pechlivanis, Nancy L. Pedersen, Leena Peltonen, John R.B. Perry, Annette Peters, Janani M. Pinidiyapathirage, Carl G.P. Platou, Simon Potter, Jackie F. Price, Lu Qi, Venkatesan Radha, Loukianos Rallidis, Asif Rasheed, Wolfgang Rathmann, Rainer Rauramaa, Soumya Raychaudhuri, N. William Rayner, Simon D. Rees, Emil Rehnberg, Samuli Ripatti, Neil Robertson, Michael Roden, Elizabeth J. Rossin, Igor Rudan, Denis Rybin, Timo E. Saaristo, Veikko Salomaa, Juha Saltevo, Maria Samuel, Dharambir KSanghera, Jouko Saramies, James Scott, Laura J. Scott, Robert A. Scott,

Ayellet V. Segrè, Joban Sehmi, Bengt Sennblad, Nabi Shah, Sonia Shah, A. Samad Shera, Xiao Ou Shu, Alan R. Shuldiner, Gunnar Sigurosson, Eric Sijbrands, Angela Silveira, Xueling Sim, Suthesh Sivapalaratnam, Kerrin S. Small, Wing Yee So, Alena Stančáková, Kari Stefansson, Gerald Steinbach, Valgerdur Steinthorsdottir, Kathleen Stirrups, Rona J. Strawbridge, Heather M. Stringham, Qi Sun, Chen Suo, Ann Christine Syvänen, Ryoichi Takayanagi, Fumihiko Takeuchi, Wan Ting Tay, Tanya M. Teslovich, Barbara Thorand, Gudmar Thorleifsson, Unnur Thorsteinsdottir, Emmi Tikkanen, Joseph Trakalo, Elena Tremoli, Mieke D. Trip, Fuu Jen Tsai, Tiinamaija Tuomi, Jaakko Tuomilehto, Andre G. Uitterlinden, Adan Valladares-Salgado, Sailaja Vedantam, Fabrizio Veglia, Benjamin F. Voight, Congrong Wang, Nicholas J. Wareham, Roman Wennauer, Ananda R. Wickremasinghe, Tom Wilsgaard, James F. Wilson, Steven Wiltshire, Wendy Winckler, Tien Yin Wong, Andrew R. Wood, Jer Yuarn Wu, Ying Wu, Ken Yamamoto, Toshimasa Yamauchi, Mingyu Yang, Loic Yengo, Mitsuhiro Yokota, Robin Young, Delilah Zabaneh, Fan Zhang, Rong Zhang, Wei Zheng, Paul Z. Zimmet, David Altshuler, Donald W. Bowden, Yoon Shin Cho, Nancy J. Cox, Miguel Cruz, Craig L. Hanis, Jaspal Kooner, Jong Young Lee, Mark Seielstad, Yik Ying Teo, Michael Boehnke, Esteban J. Parra, John C. Chambers, E. Shyong Tai, Mark I. McCarthy, and Andrew P. Morris. Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. *Nature Genetics*, 46(3), 2014.

[13] Arushi Varshney, Hadley Vanrenterghem, Peter Orchard, Alan P. Boyle, Michael L. Stitzel, Duygu Ucar, and Stephen C.J. Parker. Cell specificity of human regulatory annotations and their genetic effects on gene expression. *Genetics*, 211(2), 2019.

[14] Jacklyn N. Hellwege, Jacob M. Keaton, Ayush Giri, Xiaoyi Gao, Digna R. Velez Edwards, and Todd L. Edwards. Population Stratification in Genetic Association Studies. *Current Protocols in Human Genetics*, 95(1), 2017.

[15] Bettina Mieth, Marius Kloft, Juan Antonio Rodríguez, Sören Sonnenburg, Robin Vobruba, Carlos Morcillo-Suárez, Xavier Farré, Urko M. Marigorta, Ernst Fehr, Thorsten Dickhaus, Gilles Blanchard, Daniel Schunk, Arcadi Navarro, and Klaus Robert Müller. Combining multiple hypothesis testing with machine learning increases the statistical power of genome-wide association studies. *Scientific Reports*, 6, 2016.

[16] Lucia A. Hindorff, Praveen Sethupathy, Heather A. Junkins, Erin M. Ramos, Jayashri P. Mehta, Francis S. Collins, and Teri A. Manolio. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America*, 106(23), 2009.

[17] Anubha Mahajan, Jennifer Wessel, Sara M. Willems, Wei Zhao, Neil R. Robertson, Audrey Y. Chu, Wei Gan, Hidetoshi Kitajima, Daniel Taliun, N. William Rayner, Xiuqing Guo, Yingchang Lu, Man Li, Richard A. Jensen, Yao Hu, Shaofeng Huo, Kurt K. Lohman, Weihua Zhang, James P. Cook, Bram Peter Prins, Jason Flannick, Niels Grarup, Vassily Vladimirovich Trubetskoy, Jasmina Kravic, Young Jin Kim, Denis V. Rybin, Hanieh Yaghootkar, Martina Müller-Nurasyid, Karina Meidtner, Ruifang Li-Gao, Tibor V. Varga, Jonathan Marten, Jin Li, Albert Vernon Smith, Ping An, Symen Ligthart, Stefan Gustafsson, Giovanni Malerba, Ayse Demirkan, Juan Fernandez Tajes, Valgerdur Steinthorsdottir, Matthias Wuttke, Cécile Lecoeur, Michael Preuss, Lawrence F. Bielak, Marielisa Graff, Heather M. Highland, Anne E. Justice, Dajiang J. Liu, Eirini Marouli, Gina Marie Peloso, Helen R. Warren, Saima Afaq, Shoaib Afzal, Emma Ahlqvist, Lia B. Bang,

Alain G. Bertoni, Cristina Bombieri, Jette Bork-Jensen, Ivan Brandslund, Jennifer A. Brody, Noël P. Burtt, Mickaël Canouil, Yii Der Ida Chen, Yoon Shin Cho, Cramer Christensen, Sophie V. Eastwood, Kai Uwe Eckardt, Krista Fischer, Giovanni Gambaro, Vilmantas Giedraitis, Megan L. Grove, Hugoline G. De Haan, Sophie Hackinger, Yang Hai, Sohee Han, Anne Tybjærg-Hansen, Marie France Hivert, Bo Isomaa, Susanne Jäger, Marit E. Jørgensen, Torben Jørgensen, Annemari Käräjämäki, Bong Jo Kim, Sung Soo Kim, Heikki A. Koistinen, Peter Kovacs, Jennifer Kriebel, Florian Kronenberg, Kristi Läll, Leslie A. Lange, Jung Jin Lee, Benjamin Lehne, Huaixing Li, Keng Hung Lin, Allan Linneberg, Ching Ti Liu, Jun Liu, Marie Loh, Reedik Mägi, Vasiliki Mamakou, Roberta McKean-Cowdin, Girish Nadkarni, Matt Neville, Sune F. Nielsen, Ioanna Ntalla, Patricia A. Peyser, Wolfgang Rathmann, Kenneth Rice, Stephen S. Rich, Line Rode, Olov Rolandsson, Sebastian Schönherr, Elizabeth Selvin, Kerrin S. Small, Alena Stančáková, Praveen Surendran, Kent D. Taylor, Tanya M. Teslovich, Barbara Thorand, Gudmar Thorleifsson, Adrienne Tin, Anke Tönjes, Anette Varbo, Daniel R. Witte, Andrew R. Wood, Pranav Yajnik, Jie Yao, Loïc Yengo, Robin Young, Heiner Boeing, Eric Boerwinkle, Erwin P. Bottinger, Rajiv Chowdhury, George Dedoussis, Abbas Dehghan, Panos Deloukas, Marco M. Ferrario, Jean Ferrières, Jose C. Florez, Philippe Frossard, Vilmundur Gudnason, Tamara B. Harris, Susan R. Heckbert, Joanna M.M. Howson, Martin Ingelsson, Sekar Kathiresan, Frank Kee, Johanna Kuusisto, Claudia Langenberg, Lenore J. Launer, Cecilia M. Lindgren, Satu Männistö, Thomas Meitinger, Karen L. Mohlke, Marie Moitry, Andrew D. Morris, Alison D. Murray, Renée De Mutsert, Marju Orho-Melander, Katharine R. Owen, Markus Perola, Annette Peters, Michael A. Province, Asif Rasheed, Paul M. Ridker, Fernando Rivadineira, Frits R. Rosendaal, Anders H. Rosengren, Veikko Salomaa, Wayne H.H. Sheu, Rob Sladek, Blair H. Smith, Konstantin Strauch, André G. Uitterlinden, Rohit Varma, Cristen J. Willer, Matthias Blüher, John Campbell Chambers, John Danesh, Cornelia Van Duijn, Josée Dupuis, Oscar H. Franco, Paul W. Franks, Philippe Froguel, Harald Grallert, Leif Groop, Bok Ghee Han, Torben Hansen, Andrew T. Hattersley, Caroline Hayward, Erik Ingelsson, Sharon L.R. Kardia, Fredrik Karpe, Jaspal Singh Kooner, Anna Köttgen, Kari Kuulasmaa, Markku Laakso, Xu Lin, Lars Lind, Yongmei Liu, Ruth J.F. Loos, Jonathan Marchini, Andres Metspalu, Dennis Mook-Kanamori, Børge G. Nordestgaard, Colin N.A. Palmer, James S. Pankow, Oluf Pedersen, Bruce M. Psaty, Rainer Rauramaa, Naveed Sattar, Matthias B. Schulze, Nicole Soranzo, Timothy D. Spector, Kari Stefansson, Michael Stumvoll, Unnur Thorsteinsdottir, Tiinamaija Tuomi, Jaakko Tuomilehto, Nicholas J. Wareham, James G. Wilson, Eleftheria Zeggini, Robert A. Scott, Inês Barroso, Timothy M. Frayling, Mark O. Goodarzi, James B. Meigs, Michael Boehnke, Danish Saleheen, Andrew P. Morris, Jerome I. Rotter, and Mark I. McCarthy. Refining the accuracy of validated target identification through coding variant fine-mapping in type 2 diabetes article. *Nature Genetics*, 50(4), 2018.

[18] Hannah L. Nicholls, Christopher R. John, David S. Watson, Patricia B. Munroe, Michael R. Barnes, and Claudia P. Cabrera. Reaching the End-Game for GWAS: Machine Learning Approaches for the Prioritization of Complex Disease Loci, 2020.

[19] Sudesna Chatterjee, Kamlesh Khunti, and Melanie J. Davies. Type 2 diabetes, 6 2017.

[20] International Diabetes Federation. *International Diabetes Federation. IDF Diabetes Atlas, 7th edn. Brussels, Belgium: International Diabetes Federation, http://www.diabetesatlas.org*. 2015.

[21] Matthew T. Maurano, Richard Humbert, Eric Rynes, Robert E. Thurman, Eric Haugen, Hao Wang, Alex P. Reynolds, Richard Sandstrom, Hongzhu Qu, Jennifer Brody, Anthony Shafer, Fidencio Neri, Kristen Lee, Tanya Kutyavin, Sandra Stehling-Sun, Audra K. Johnson, Theresa K. Canfield, Erika Giste, Morgan Diegel, Daniel Bates, R. Scott Hansen, Shane Neph, Peter J. Sabo, Shelly Heimfeld, Antony Raubitschek, Steven Ziegler, Chris Cotsapas, Nona Sotoodehnia, Ian Glass, Shamil R. Sunyaev, Rajinder Kaul, and John A. Stamatoyannopoulos. Systematic localization of common disease-associated variation in regulatory DNA. *Science*, 337(6099), 2012.

[22] Lorenzo Pasquali, Kyle J. Gaulton, Santiago A. Rodríguez-Seguí, Loris Mularoni, Irene Miguel-Escalada, Ildem Akerman, Juan J. Tena, Ignasi Morán, Carlos Gómez-Marín, Martijn Van De Bunt, Joan Ponsa-Cobas, Natalia Castro, Takao Nammo, Inês Cebola, Javier García-Hurtado, Miguel Angel Maestro, François Pattou, Lorenzo Piemonti, Thierry Berney, Anna L. Gloyn, Philippe Ravassard, José Luis Gómez Skarmeta, Ferenc Müller, Mark I. Mccarthy, and Jorge Ferrer. Pancreatic islet enhancer clusters enriched in type 2 diabetes risk-associated variants. *Nature Genetics*, 46(2), 2014.

[23] Teri A. Manolio, Francis S. Collins, Nancy J. Cox, David B. Goldstein, Lucia A. Hindorff, David J. Hunter, Mark I. McCarthy, Erin M. Ramos, Lon R. Cardon, Aravinda Chakravarti, Judy H. Cho, Alan E. Guttmacher, Augustine Kong, Leonid Kruglyak, Elaine Mardis, Charles N. Rotimi, Montgomery Slatkin, David Valle, Alice S. Whittemore, Michael Boehnke, Andrew G. Clark, Evan E. Eichler, Greg Gibson, Jonathan L. Haines, Trudy F.C. MacKay, Steven A. McCarroll, and Peter M. Visscher. Finding the missing heritability of complex diseases, 2009.

[24] Angli Xue, Yang Wu, Zhihong Zhu, Futao Zhang, Kathryn E. Kemper, Zhili Zheng, Loic Yengo, Luke R. Lloyd-Jones, Julia Sidorenko, Yeda Wu, Mawussé Agbessi, Habibul Ahsan, Isabel Alves, Anand Andiappan, Philip Awadalla, Alexis Battle, Frank Beutner, Marc Jan J. Bonder, Dorret Boomsma, Mark Christiansen, Annique Claringbould, Patrick Deelen, Tõnu Esko, Marie Julie Favé, Lude Franke, Timothy Frayling, Sina Gharib, Gregory Gibson, Gibran Hemani, Rick Jansen, Mika Kähönen, Anette Kalnapenkis, Silva Kasela, Johannes Kettunen, Yungil Kim, Holger Kirsten, Peter Kovacs, Knut Krohn, Jaanika Kronberg-Guzman, Viktorija Kukushkina, Zoltan Kutalik, Bernett Lee, Terho Lehtimäki, Markus Loeffler, Urko M. Marigorta, Andres Metspalu, Lili Milani, Martina Müller-Nurasyid, Matthias Nauck, Michel Nivard, Brenda Penninx, Markus Perola, Natalia Pervjakova, Brandon Pierce, Joseph Powell, Holger Prokisch, Bruce Psaty, Olli Raitakari, Susan Ring, Samuli Ripatti, Olaf Rotzschke, Sina Ruëger, Ashis Saha, Markus Scholz, Katharina Schramm, Ilkka Seppälä, Michael Stumvoll, Patrick Sullivan, Alexander Teumer, Joachim Thiery, Lin Tong, Anke Tönjes, Jenny van Dongen, Joyce van Meurs, Joost Verlouw, Uwe Völker, Urmo Võsa, Hanieh Yaghootkar, Biao Zeng, Allan F. McRae, Peter M. Visscher, Jian Zeng, and Jian Yang. Genome-wide association analyses identify 143 risk variants and putative regulatory mechanisms for type 2 diabetes. *Nature Communications*, 9(1), 12 2018.

[25] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *Springer Series in Statistics The Elements of Statistical Learning - Data Mining, Inference, and Prediction*, volume 2nd. 2009.

[26] Gareth James, Daniela Witten, and Trevor Hastie. *Introduction to Statistical Learning with Applications in R*, volume 11. 2019.

[27] Rene Y. Choi, Aaron S. Coyner, Jayashree Kalpathy-Cramer, Michael F. Chiang, and J. Peter Campbell. Introduction to machine learning, neural networks, and deep learning. *Translational Vision Science and Technology*, 9(2), 2020.

[28] S. B. Kotsiantis. Supervised machine learning: A review of classification techniques, 2007.

[29] Zoubin Ghahramani. Unsupervised Learning BT - Advanced Lectures on Machine Learning. *Advanced Lectures on Machine Learning*, 3176(Chapter 5), 2004.

[30] Isha Salian. Difference Between Supervised, Unsupervised, & Reinforcement Learning — NVIDIA Blog, 2018.

[31] D. O. HEBB. Animal and physiological psychology. *Annual review of psychology*, 1, 1950.

[32] Bengio Y. Hinton G. LeCun, Y. Deep learning. nature 521 (7553): 436. *Nature*, 521, 2015.

[33] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088), 1986.

[34] B. W. White and Frank Rosenblatt. Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms. *The American Journal of Psychology*, 76(4), 1963.

[35] J. Leonard and M. A. Kramer. Improvement of the backpropagation algorithm for training neural networks. *Computers and Chemical Engineering*, 14(3), 1990.

[36] Crescenzio Gallo. Artificial Neural Networks Tutorial. In *Encyclopedia of Information Science and Technology, Third Edition*. 2014.

[37] Xinghuo Yu, M. Onder Efe, and Okyay Kaynak. A general backpropagation algorithm for feedforward neural networks learning, 2002.

[38] Ilya Tolstikhin, Olivier Bousquet, Bernhard Schölkopf, Konstantin Thierbach, Pierre Louis Bazin, Walter de Back, Filippos Gavriilidis, Evgeniya Kirilina, Carsten Jäger, Markus Morawski, Stefan Geyer, Nikolaus Weiskopf, Nico Scherf, Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio, Elon Musk, Neuralink, Martin A Hjortsø, Peter Wolenski, Sebastian Ruder, Will Grathwohl, Ricky T. Q. Chen, Jesse Bettencourt, Ilya Sutskever, David Duvenaud, and Carl Doersch. An overview of gradient descent optimization algorithms. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11046 LNCS(NeurIPS), 2018.

[39] Nathan Killoran, Leo J. Lee, Andrew Delong, David Duvenaud, and Brendan J. Frey. Generating and designing DNA with deep generative models, 2017.

[40] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Networks. 6 2014.

[41] Liis Leitsalu, Toomas Haller, Tõnu Esko, Mari Liis Tammesoo, Helene Alavere, Harold Snieder, Markus Perola, Pauline C. Ng, Reedik Mägi, Lili Milani, Krista Fischer, and Andres Metspalu. Cohort profile: Estonian biobank of the Estonian genome center, university of Tartu. *International Journal of Epidemiology*, 44(4), 2015.

[42] H. M. Cann. A Human Genome Diversity Cell Line Panel. *Science*, 296(5566), 2002.

[43] Swapan Mallick, Heng Li, Mark Lipson, Iain Mathieson, Melissa Gymrek, Fernando Racimo, Mengyao Zhao, Niru Chennagiri, Susanne Nordenfelt, Arti Tandon, Pontus Skoglund, Iosif Lazaridis, Sriram Sankararaman, Qiaomei Fu, Nadin Rohland, Gabriel Renaud, Yaniv Erlich, Thomas Willems, Carla Gallo, Jeffrey P. Spence, Yun S. Song, Giovanni Poletti, Francois Balloux, George Van Driem, Peter De Knijff, Irene Gallego Romero, Aashish R. Jha, Doron M. Behar, Claudio M. Bravi, Cristian Capelli, Tor Hervig, Andres Moreno-Estrada, Olga L. Posukh, Elena Balanovska, Oleg Balanovsky, Sena Karachanak-Yankova, Hovhannes Sahakyan, Draga Toncheva, Levon Yepiskoposyan, Chris Tyler-Smith, Yali Xue, M. Syafiq Abdullah, Andres Ruiz-Linares, Cynthia M. Beall, Anna Di Rienzo, Choongwon Jeong, Elena B. Starikovskaya, Ene Metspalu, Jüri Parik, Richard Villems, Brenna M. Henn, Ugur Hodoglugil, Robert Mahley, Antti Sajantila, George Stamatoyannopoulos, Joseph T.S. Wee, Rita Khusainova, Elza Khusnutdinova, Sergey Litvinov, George Ayodo, David Comas, Michael F. Hammer, Toomas Kivisild, William Klitz, Cheryl A. Winkler, Damian Labuda, Michael Bamshad, Lynn B. Jorde, Sarah A. Tishkoff, W. Scott Watkins, Mait Metspalu, Stanislav Dryomov, Rem Sukernik, Lalji Singh, Kumarasamy Thangaraj, Svante Paäbo, Janet Kelso, Nick Patterson, and David Reich. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature*, 538(7624), 2016.

[44] Alice B. Popejoy and Stephanie M. Fullerton. Genomics is failing on diversity, 2016.

[45] Giorgio Sirugo, Scott M. Williams, and Sarah A. Tishkoff. The Missing Diversity in Human Genetic Studies, 2019.

[46] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *Advances in Neural Information Processing Systems*, 2016.

[47] Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert E. Handsaker, Gerton Lunter, Gabor T. Marth, Stephen T. Sherry, Gilean McVean, and Richard Durbin. The variant call format and VCFtools. *Bioinformatics*, 27(15), 2011.

[48] Christopher C. Chang, Carson C. Chow, Laurent C.A.M. Tellier, Shashaank Vattikuti, Shaun M. Purcell, and James J. Lee. Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience*, 4(1), 2015.

[49] François Chollet. Keras: The Python Deep Learning library. *Keras.Io*, 2015.

[50] W. McKinney. Data Structures for Statistical Computing in Python. In Proceedings of the 9th Python in Science Conference. In *Proceedings of the 9th Python in Science Conference*, 2010.

[51] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith,

Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy, 2020.

[52] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2011.

[53] Raphael Vallat. Pingouin: statistics in Python. *Journal of Open Source Software*, 3(31), 2018.

[54] Michael Waskom. seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60), 2021.

[55] Daniel Savic, Honggang Ye, Ivy Aneas, Soo Young Park, Graeme I. Bell, and Marcelo A. Nobrega. Alterations in TCF7L2 expression define its role as a key regulator of glucose metabolism. *Genome Research*, 21(9), 2011.

[56] Ludmila Prokunina-Olsson, Cullan Welch, Ola Hansson, Neeta Adhikari, Laura J. Scott, Nicolle Usher, Maurine Tong, Andrew Sprau, Amy Swift, Lori L. Bonnycastle, Michael R. Erdos, Zhi He, Richa Saxena, Brennan Harmon, Olga Kotova, Eric P. Hoffman, David Altshuler, Leif Groop, Michael Boehnke, Francis S. Collins, and Jennifer L. Hall. Tissue-specific alternative splicing of TCF7L2. *Human Molecular Genetics*, 18(20), 2009.

[57] Melissa J. Landrum, Jennifer M. Lee, Mark Benson, Garth R. Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, Jennifer Hart, Douglas Hoffman, Wonhee Jang, Karen Karapetyan, Kenneth Katz, Chunlei Liu, Zenith Maddipatla, Adriana Malheiro, Kurt McDaniel, Michael Ovetsky, George Riley, George Zhou, J. Bradley Holmes, Brandi L. Kattman, and Donna R. Maglott. ClinVar: Improving access to variant interpretations and supporting evidence. *Nucleic Acids Research*, 46(D1), 2018.

[58] Tomaz Berisa and Joseph K. Pickrell. Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics*, 32(2), 2016.

[59] Anthony L. Hinrichs, Emma K. Larkin, and Brian K. Suarez. Population stratification and patterns of linkage disequilibrium. In *Genetic Epidemiology*, volume 33, 2009.

[60] Gift Nyamundanda, Pawan Poudel, Yatish Patil, and Anguraj Sadanandam. A Novel Statistical Method to Diagnose, Quantify and Correct Batch Effects in Genomic Studies. *Scientific Reports*, 7(1), 12 2017.

[61] Matthieu Bouaziz, Christophe Ambroise, and Mickael Guedj. Accounting for population stratification in practice: A comparison of the main strategies dedicated to genome-wide association studies. *PLoS ONE*, 6(12), 2011.

[62] Kosuke Hamazaki and Hiroyoshi Iwata. Rainbow: Haplotype-based genome-wide association study using a novel SNP-set method. *PLoS Computational Biology*, 16(2), 2020.

[63] Yaël Frégier and Jean Baptiste Gouray. Mind2Mind: transfer learning for GANs, 2019.

[64] Mehdi Mirza and Simon Osindero. CGAN. *CoRR*, 2014.

[65] David W. Craig, Robert M. Goor, Zhenyuan Wang, Justin Paschall, Jim Ostell, Michael Feolo, Stephen T. Sherry, and Teri A. Manolio. Assessing and managing risk when sharing aggregate genetic variant data, 2011.

Appendices

A Python chopping script

```
import pandas as pd


inpt = '/Users/kovgl/artificial_genomes/chr7_t2d_filtered_control.hapt'

#Read input
df = pd.read_csv(inpt, sep = ' ', header=None)
#df = df.sample(frac=1).reset_index(drop=True)
df_noname = df.drop(df.columns[0:6], axis=1)

#Check number of columns (SNPs)
num_of_snps = len(df_noname.columns)
snps = num_of_snps

#Divide dataframe (1,000 SNPs - one piece)
count = 0
step = 1000 #Divide by...
cols = 0
shift = step

while True:

    count = count + 1

    if (snps - step) > 0:
        splitted_genomes_df = df_noname.iloc[:,cols:shift]
        #print(df1.shape)
        splitted_genomes_df.insert(loc=0, column=str(df.columns[1]),
        value=df.iloc[:,1].values)
        splitted_genomes_df.insert(loc=0, column=str(df.columns[0]),
        value=df.iloc[:,0].values)
```

```
        #Output in hapt format
        splitted_genomes_df.to_csv(
        "chr7_control_"+str(count)+"_splitted.hapt",
        sep=" ", header=False, index=False)
        cols = cols + step
        shift = shift + step
        snps = snps - step

    elif (snps - step) == 0:
        splitted_genomes_df = df_noname.iloc[:,cols:shift]
        #print(df1.shape)
        splitted_genomes_df.insert(loc=0, column=str(df.columns[1]),
        value=df.iloc[:,1].values)
        splitted_genomes_df.insert(loc=0, column=str(df.columns[0]),
        value=df.iloc[:,0].values)
        #Output in hapt format
        splitted_genomes_df.to_csv(
        "chr7_control_"+str(count)+"_splitted.hapt",
        sep=" ", header=False, index=False)
        cols = cols + step
        shift = shift + step
        snps = snps - step
        break

    else: #negative
        shift = cols + (num_of_snps - cols)
        print("Last shift is: " + str(shift))
        splitted_genomes_df = df_noname.iloc[:,cols:shift]
        #print(df2.shape)
        splitted_genomes_df.insert(loc=0, column=str(df.columns[1]),
        value=df.iloc[:,1].values)
        splitted_genomes_df.insert(loc=0, column=str(df.columns[0]),
        value=df.iloc[:,0].values)
        #Output in hapt format
        splitted_genomes_df.to_csv(
        "chr7_control_"+str(count)+"_splitted.hapt",
        sep=" ", header=False, index=False)
        break
```

B Python stitching script

```
import pandas as pd

def readf(inpt):
    global df
    #Read input
    df = pd.read_csv(inpt, sep = ' ', header=None)
    #df = df.sample(frac=1).reset_index(drop=True)
    df_noname = df.drop(df.columns[0:2], axis=1)
    #df_noname = df_noname.values
    return df_noname

chrom = 7
chunks = 17

df = []

frames = [ readf("6000master_chr" +
        str(chrom) + "_control_"+str(i + 1) + "_output.hapt")
for i in range(chunks)
]
result = pd.concat(frames, axis = 1)

result.insert(loc=0, column=str(df.columns[1]),
value=df.iloc[:,1].values)
result.insert(loc=0, column=str(df.columns[0]),
value=df.iloc[:,0].values)
#Output in hapt format
result.to_csv("chr"+str(chrom)+"_AG_control_stitched.hapt", sep=" ",
header=False, index=False)
```

C Chromosome 22, chunk 1

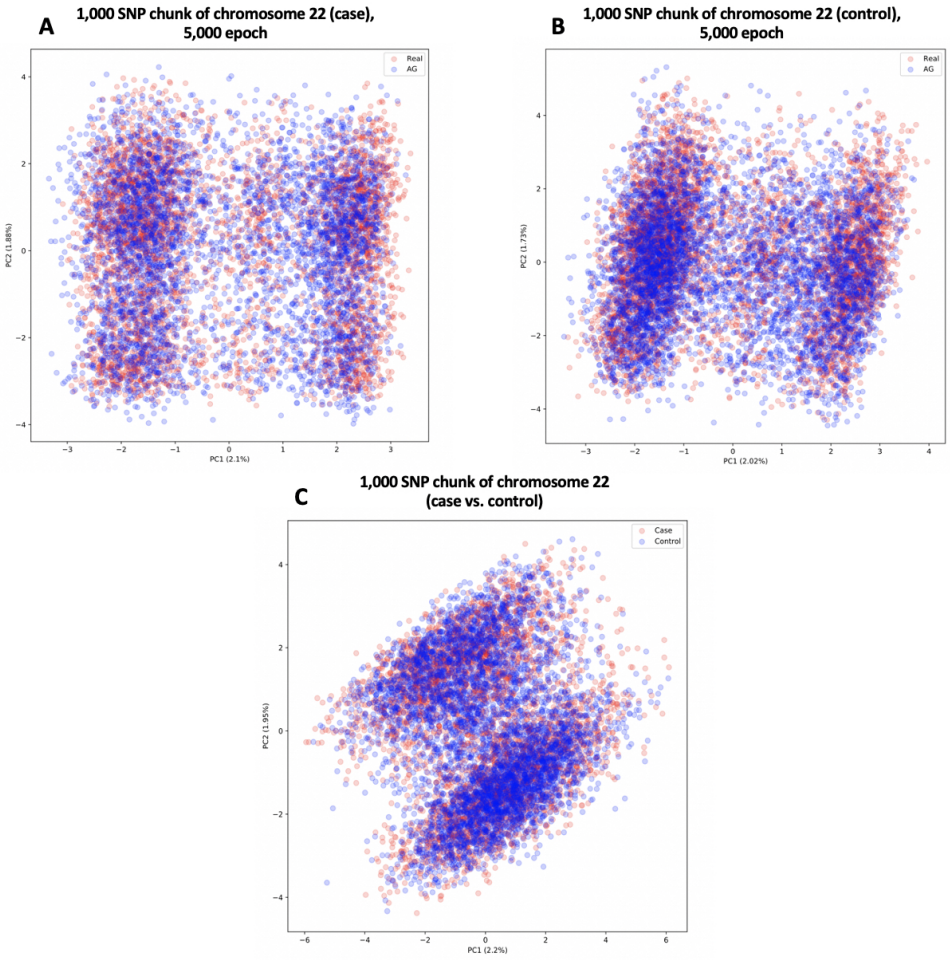

Figure C: PCA analysis of 1,000 SNP chunks (chromosome 22, chunk 1). The X-axis represents the first principal component (PC1), the Y-axis represents the second principal component (PC2). For subplots A-B, real genomes are displayed in red color, while artificial genomes (AG) are displayed in blue color. For subplot C, case artificial genomes are displayed in red color, while control artificial genomes are displayed in blue color. (A) Chromosome 22 (case) first chunk training at 5,000 epoch. (B) Chromosome 22 (control) first chunk training at 5,000 epoch. (C) Chromosome 22 first chunk artificial genomes cases with controls.

D Chromosome 7, chunk 13

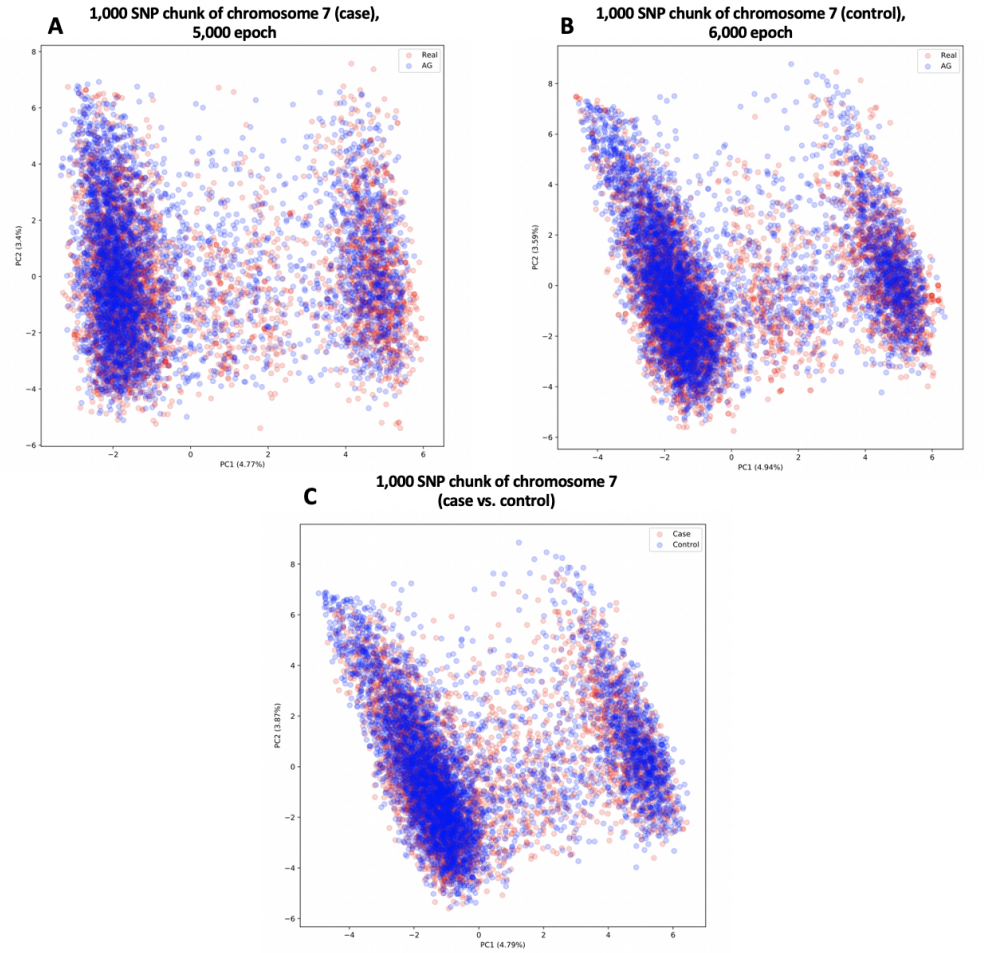

Figure D: PCA analysis of 1,000 SNP chunks (chromosome 7, chunk 13). The X-axis represents the first principal component (PC1), the Y-axis represents the second principal component (PC2). For subplots A-B, real genomes are displayed in red color, while artificial genomes (AG) are displayed in blue color. For subplot C, case artificial genomes are displayed in red color, while control artificial genomes are displayed in blue color. (A) Chromosome 7 (case) thirteenth chunk training at 5,000 epoch. (B) Chromosome 7 (control) thirteenth chunk training at 6,000 epoch. (C) Chromosome 7 thirteenth chunk artificial genomes cases with controls.

E Chromosome 22

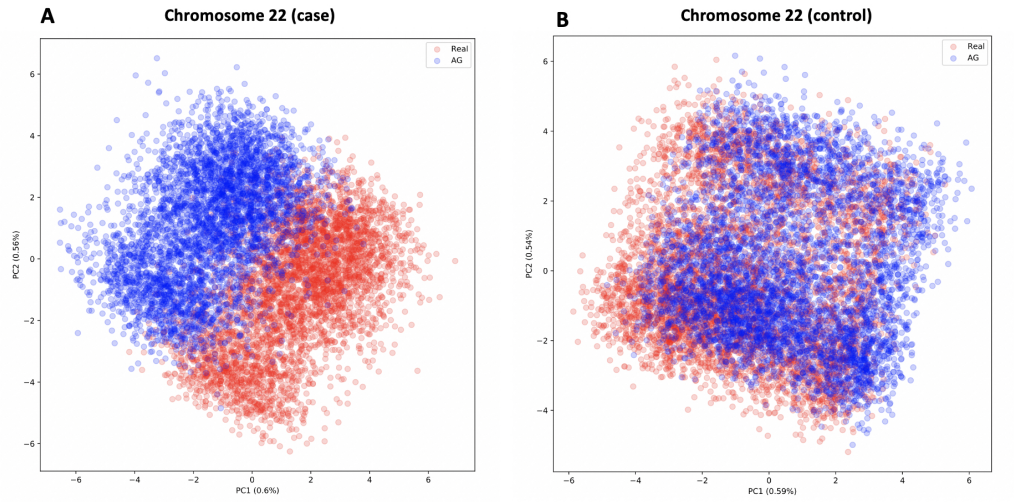

Figure E: PCA analysis of real genomes with stitched artificial genomes (chromosome 22). The X-axis represents the first principal component (PC1), the Y-axis represents the second principal component (PC2). The percentage of variance explained by each principal component is displayed on the axis labels. Real genomes are displayed in red color, while artificial genomes (AG) are displayed in blue color. (A) Chromosome 22 (case). (B) Chromosome 22 (control).

F Chromosome 7

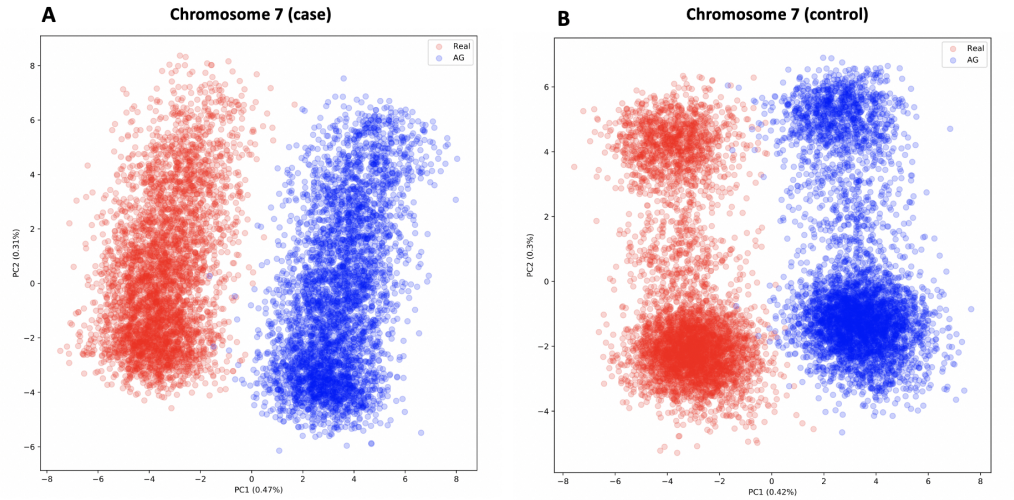

Figure F: PCA analysis of real genomes with stitched artificial genomes (chromosome 7). The X-axis represents the first principal component (PC1), the Y-axis represents the second principal component (PC2). The percentage of variance explained by each principal component is displayed on the axis labels. Real genomes are displayed in red color, while artificial genomes (AG) are displayed in blue color. (A) Chromosome 7 (case). (B) Chromosome 7 (control).

G Chromosome 22 (Case vs. Control)

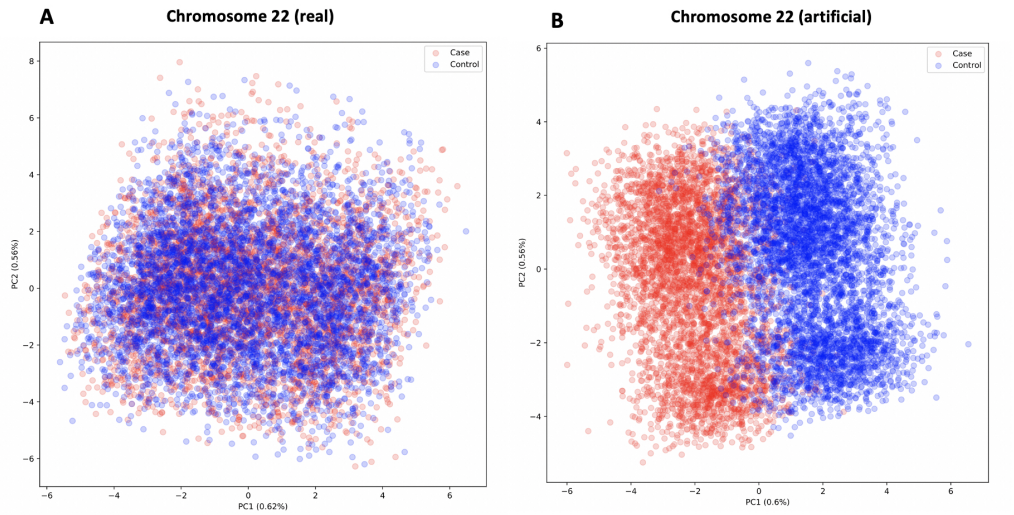

Figure G: PCA analysis of cases with controls (chromosome 22). The X-axis represents the first principal component (PC1), the Y-axis represents the second principal component (PC2). The percentage of variance explained by each principal component is displayed on the axis labels. Cases are displayed in red color, while controls are displayed in blue color. (A) Chromosome 22 real genomes. (B) Chromosome 22 artificial genomes.

H Chromosome 7 (Case vs. Control)

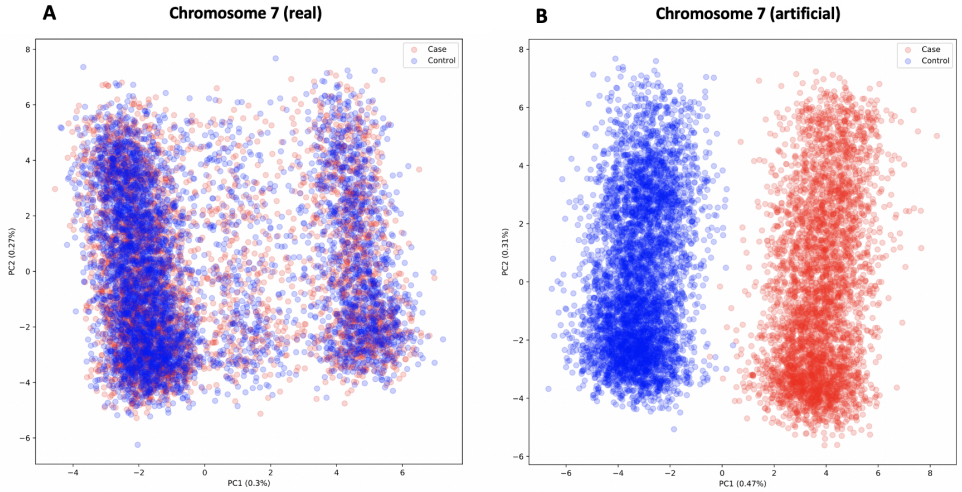

Figure H: PCA analysis of cases with controls (chromosome 7). The X-axis represents the first principal component (PC1), the Y-axis represents the second principal component (PC2). The percentage of variance explained by each principal component is displayed on the axis labels. Cases are displayed in red color, while controls are displayed in blue color. (A) Chromosome 7 real genomes. (B) Chromosome 7 artificial genomes.

I A systematic error accumulation between AGs cases and controls

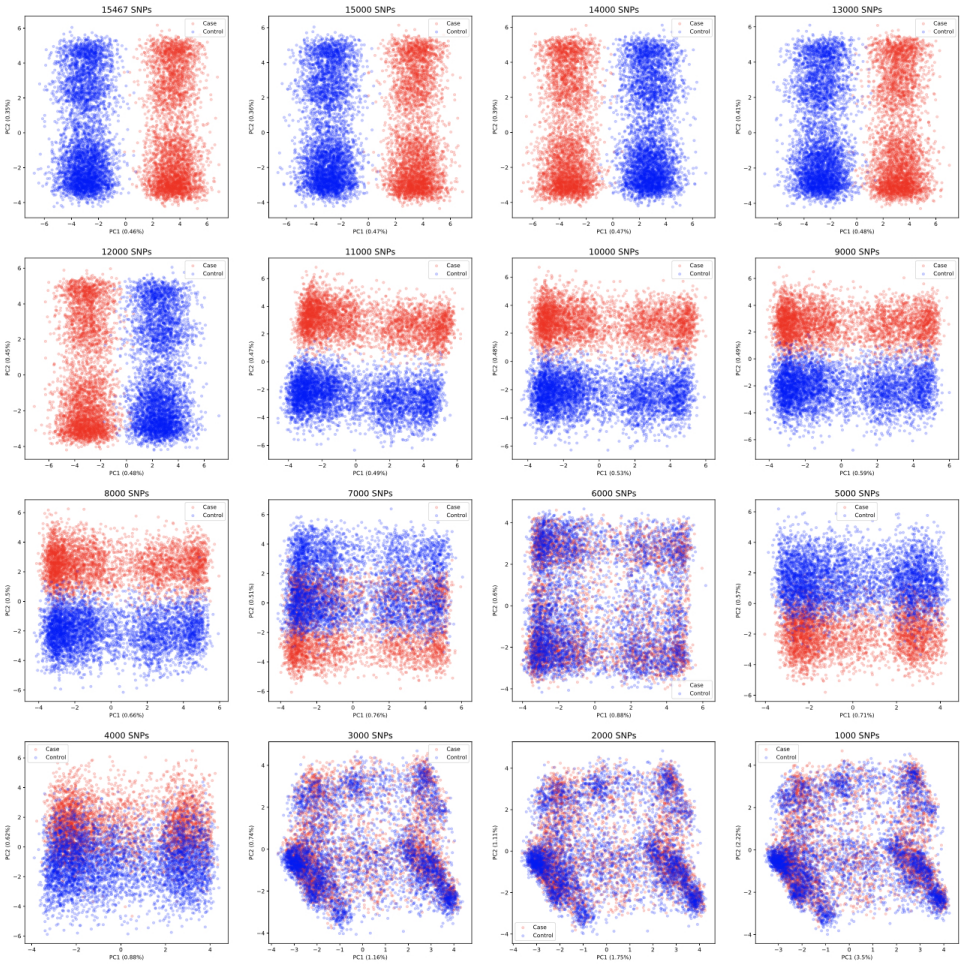

Figure I: Demonstration of a systematic difference accumulation between artificial genomes cases and controls during the stitching process. The analysis was performed on the artificial chromosome 10. Subplots represent PCA of different number of SNPs coming from different number of chunks stitched together. First subplot with 15,467 SNPs represents the full chromosome, while the last subplot with 1,000 SNPs represents one (first) training chunk. On each subplot, the X-axis represents the first principal component (PC1), the Y-axis represents the second principal component (PC2). The percentage of variance explained by each principal component is displayed on the axis labels. Cases are displayed in red color, while controls are displayed in blue color.

J MAF correlation analysis (chromosome 7)

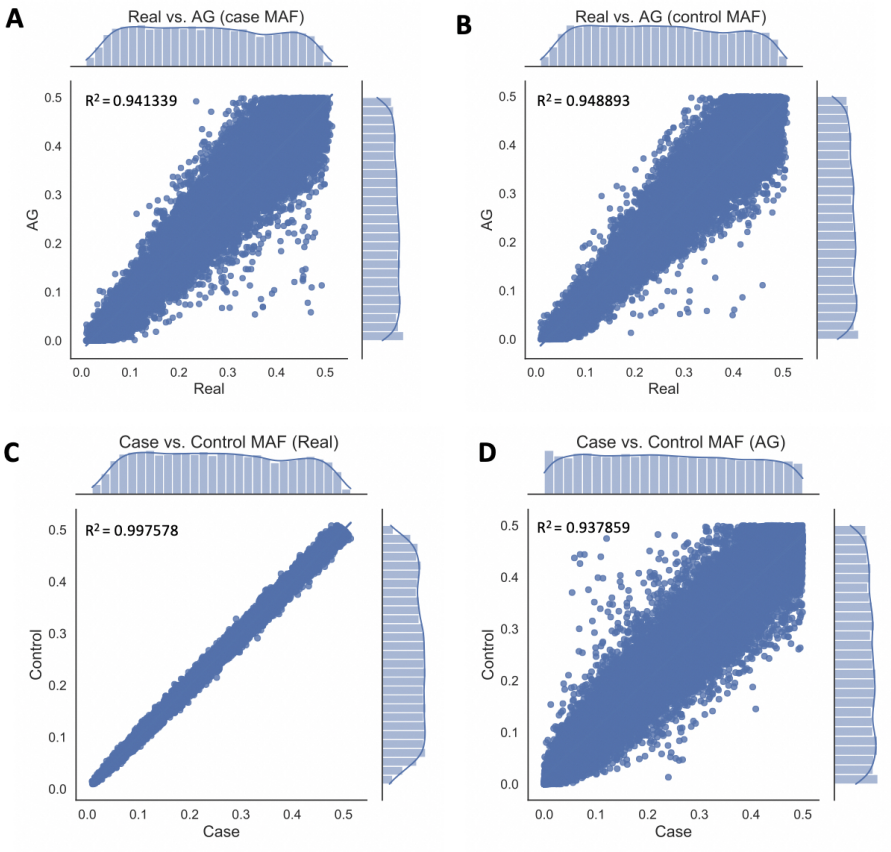

Figure J: Minor allele frequency (MAF) correlation analysis between real and artificial genomes (chromosome 7). (A) Case MAF correlation between real and artificial genomes. (B) Control MAF correlation between real and artificial genomes. (C) Real genomes MAF correlation between cases and controls. (D) Artificial genomes (AG) MAF correlation between cases and controls. The X-axis represents MAF of real genomes (A, B) or cases (C, D), while the Y-axis represents MAF of artificial genomes (AG) (A, B) or controls (C,D). The marginal distributions are indicated as histograms on the sides. The coefficient of determination (R2) is also displayed on the plots.

K MAF correlation analysis (chromosome 7, chunk 13)

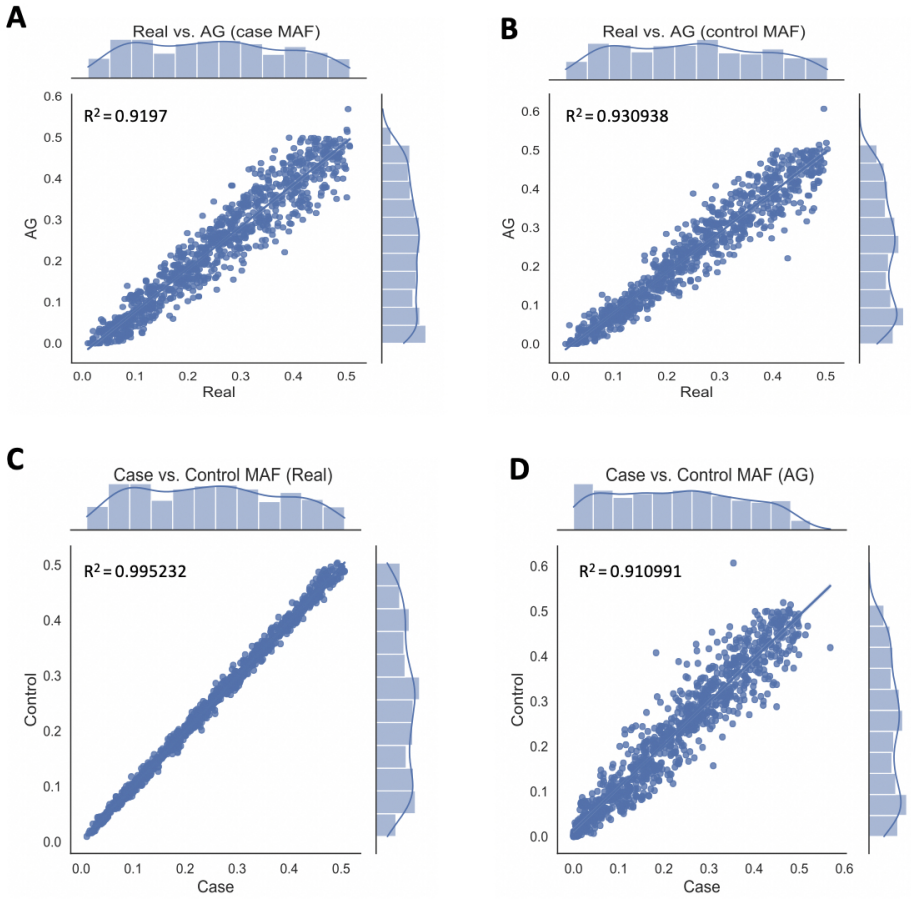

Figure K: Minor allele frequency (MAF) correlation analysis between real and artificial genomes (chromosome 7, chunk 13). (A) Case MAF correlation between real and artificial genomes. (B) Control MAF correlation between real and artificial genomes. (C) Real genomes MAF correlation between cases and controls. (D) Artificial genomes (AG) MAF correlation between cases and controls. The X-axis represents MAF of real genomes (A, B) or cases (C, D), while the Y-axis represents MAF of artificial genomes (AG) (A, B) or controls (C,D). The marginal distributions are indicated as histograms on the sides. The coefficient of determination (R2) is also displayed on the plots.

L Artificial chromosome 7 GWAS

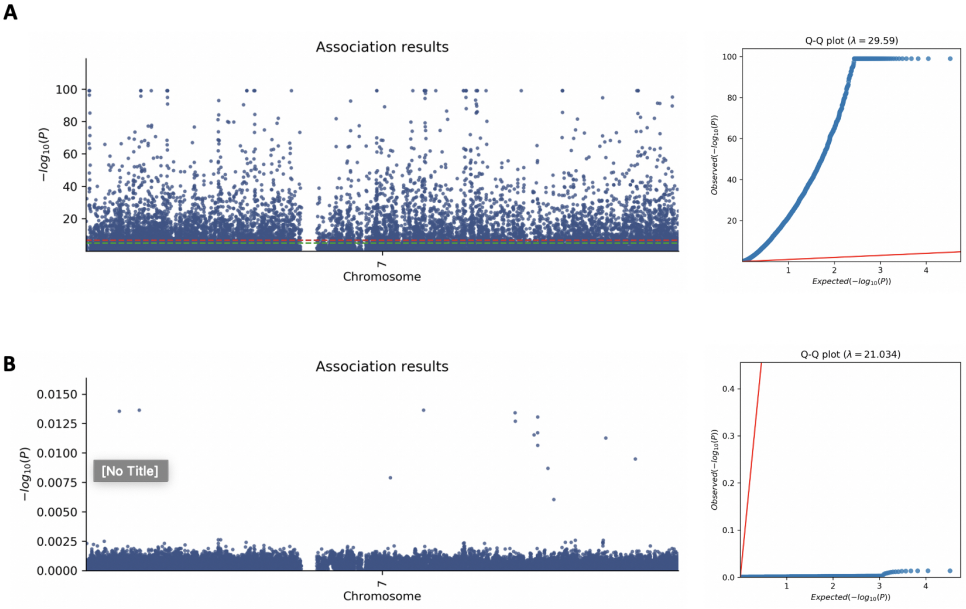

Figure L: GWAS analysis of artificial genomes chromosome 7 data. Results are represented in the form of manhattan plot. The X-axis on the figure indicates haplotypes from each tested region of the chromosome 10. The Y-axis indicates p-values in the scale of negative common logarithm. Green line represents standard suggestive threshold ($P<5 \times 10^{-8}$), while red line represents more stringent, adjusted for Bonferroni correction, genome-wide threshold ($P<6.79588 \times 10^{-8}$), creating a statistical significance borderline. Top hits crossing the suggestive threshold are marked on the plot. Results are supplemented with quantile-quantile (Q-Q) plots, where genomic inflation factor (λ) is displayed in the title. (A) Association analysis. (B) Logistic regression analysis with top 10 multidimensional scaling (MDS) components as covariates.

M Artificial chromosome 7, chunk 13 GWAS

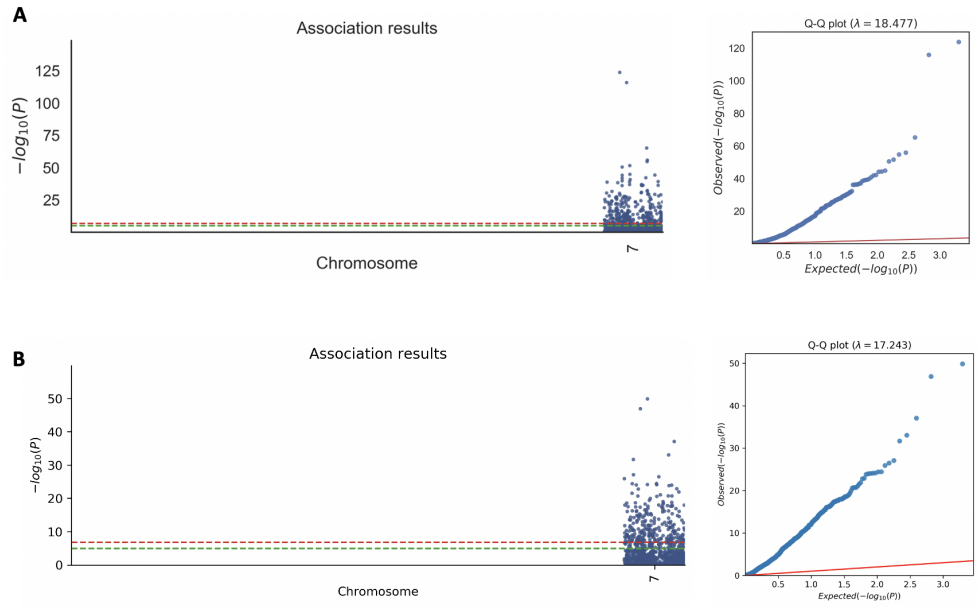

Figure M: GWAS analysis of artificial genomes chromosome 7, chunk 13 data. Results are represented in the form of manhattan plot. The X-axis on the figure indicates haplotypes from each tested region of the chunk 13. The Y-axis indicates p-values in the scale of negative common logarithm. Green line represents standard suggestive threshold ($P<5 \times 10^{-8}$), while red line represents more stringent, adjusted for Bonferroni correction, genome-wide threshold ($P<6.79588 \times 10^{-8}$), creating a statistical significance borderline. Top hits crossing the suggestive threshold are marked on the plot. Results are supplemented with quantile-quantile (Q-Q) plots, where genomic inflation factor (λ) is displayed in the title. (A) Association analysis. (B) Logistic regression analysis with top 10 multidimensional scaling (MDS) components as covariates.

Non-exclusive licence to reproduce thesis and make thesis public

I, Gleb Kovalev,

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to:

   1.1. reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright, and

   1.2. make available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright,

   "Potential of Artificial Genomes in Genome-wide Association Studies",

   supervised by Burak Yelmen and Kadir Aktas.

2. I am aware of the fact that the author retains the rights specified in p. 1.

3. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Gleb Kovalev
20.05.2021